

Il problema del contenuto semantico nella codifica predittiva gerarchica

Per una teleosemantica computazionale sincronica

I — L'intenzionalità come scandalo: genesi e struttura di un problema irrisolto

I.1 — Che cosa significa che uno stato mentale è *di* qualcosa

Tra i tratti che rendono la mente un oggetto di indagine refrattario alle categorie delle scienze naturali mature, nessuno è più persistente e più profondo di quello che Franz Brentano identificò, nella sua *Psychologie vom empirischen Standpunkt* del 1874, con il termine di intenzionalità. La mente non è semplicemente un sistema che elabora segnali: è un sistema i cui stati sono caratterizzati da una peculiare relazione di direzionalità verso oggetti che possono essere assenti, inesistenti, ipotetici o anche contraddittori. Quando percepisco una mela, il mio stato percettivo è *di* quella mela; quando ricordo un pomeriggio d'infanzia, il ricordo è *di* quell'evento passato; quando temo che piova domani, la paura è *di* un evento futuro che potrebbe non verificarsi. Questa relazione — l'*aboutness* degli stati mentali, per adottare la terminologia anglosassone divenuta quasi tecnica — non ha corrispettivo ovvio nella fisica o nella chimica. Un atomo di carbonio non è *di* nulla; una reazione enzimatica non riguarda nulla; una scarica elettrica in un conduttore non ha oggetto. Gli stati mentali sì, e questo li rende peculiari in un modo che ogni teoria della mente deve affrontare esplicitamente.

Brentano distingueva l'esistenza intenzionale degli oggetti — il modo in cui un oggetto esiste *come correlato* di uno stato mentale, indipendentemente dalla sua esistenza reale — dall'esistenza reale degli stessi. Il triangolo che immagino esiste intenzionalmente come oggetto del mio immaginare, anche se non esiste alcun triangolo fisico con le proporzioni che mi raffiguro. Questa distinzione è filosoficamente feconda ma genera immediatamente una difficoltà: se l'esistenza intenzionale è irriducibile all'esistenza fisica, come si inserisce in un quadro naturalistico dell'essere umano? E se invece l'intenzionalità è riducibile a relazioni naturali — causali, informazionali, funzionali — in che senso si rendono conto del fatto che il contenuto mentale può essere *erroneo*, può riguardare il non-esistente, può fraintendere la realtà?

È importante sottolineare, prima di procedere, che il problema dell'intenzionalità non è un problema soltanto moderno o un artefatto della filosofia analitica del Novecento. Le sue radici affondano in dispute medievali sulla natura dei concetti universali, nella scolastica di Duns Scoto e di Guglielmo di Occam, e risalgono fino all'Aristotele del *De Anima*, dove la forma dell'oggetto percepito viene in qualche senso «ricevuta» dall'organo sensorio senza la materia. L'innovazione di Brentano consiste nell'aver isolato il fenomeno con precisione terminologica e nell'averlo collocato al centro di una psicologia descrittiva sistematica, ma il problema era implicito in ogni discussione sulla natura della conoscenza e della rappresentazione. Ciò che la filosofia analitica ha aggiunto è la sfida di rendere conto dell'intenzionalità entro i vincoli del naturalismo scientifico — l'esigenza, cioè, che ogni fenomeno mentale sia descrivibile in termini compatibili con la visione del mondo delle scienze naturali.

La questione diventa ancora più acuta quando si considera la ricchezza strutturale dell'intenzionalità. Non solo gli stati mentali sono *di* qualcosa: sono di qualcosa *sotto una certa descrizione, in un certo modo di presentazione*, con una determinata *prospettiva*. Se penso a Venere come Stella del Mattino, il mio pensiero è *di* Venere sotto la descrizione «stella del mattino», e la sostituzione di questa descrizione con la sinonima «stella della sera» — che denota lo stesso pianeta — può cambiare le proprietà inferenziali del mio pensiero. Questo fenomeno, noto come opacità referenziale o intensionalità con la esse, era già tematizzato da Frege nella distinzione tra Sinn (senso) e Bedeutung (riferimento): due espressioni con lo stesso riferimento possono avere sensi diversi, e i contesti intensionali — come quelli creati dai verbi mentali «crede che», «desidera che», «teme che» — non permettono la sostituzione delle co-referenti salva veritate. Il fatto che «Lois Lane crede che Superman vola» possa essere vero mentre «Lois Lane crede che Clark Kent vola» sia falso, nonostante Superman e Clark Kent siano la stessa persona, mostra che i contesti mentali non sono estensionali: ciò che conta non è soltanto il riferimento delle espressioni ma il modo in cui il soggetto concepisce il riferimento.

Questa stratificazione — tra riferimento e modo di presentazione, tra oggetto intenzionale e contenuto — genera un secondo livello di complessità per qualsiasi teoria naturalistica. Non basta spiegare in che senso uno stato è di un oggetto; bisogna spiegare in che senso è di quell'oggetto *sotto quella descrizione, in quel modo di presentazione*. La dimensione del senso fregeano, che la filosofia analitica ha integrato nella teoria del contenuto attraverso le nozioni di contenuto stretto e contenuto ampio, di ruolo concettuale e di modo di presentazione computazionale, è una dimensione che qualsiasi teoria adeguata deve rendere conto. Come si vedrà, il framework della codifica predittiva ha le risorse per affrontare anche questa dimensione, attraverso la struttura dei livelli gerarchici del modello generativo: i livelli inferiori

corrispondono ai riferimenti, i livelli superiori ai modi di presentazione.

Quest'ultima questione — il problema della misrappresentazione o della rappresentazione scorretta — è il banco di prova più severo per qualsiasi teoria naturalistica del contenuto semantico. Una pietra colpisce un'altra pietra e l'urto produce un moto: questo è un processo causale, ma la pietra non sbaglia, non fraintende, non rappresenta scorrettamente. Un sistema radar rileva un oggetto metallico e lo registra come aeromobile: se l'oggetto è un uccello, il radar sbaglia. Ma in che senso sbaglia? La risposta intuitiva — ha prodotto un segnale che non corrisponde alla realtà — presuppone già la nozione di contenuto semantico che si intende spiegare: presuppone che il segnale abbia un contenuto determinato rispetto al quale si può essere in errore. La circolarità è reale e costituisce il nucleo del problema della misrappresentazione, che in inglese è stato denominato *the disjunction problem* da Fodor nella sua formulazione più precisa: se il contenuto di uno stato è determinato dalle sue condizioni causali di occorrenza, allora uno stato causato sia da anatre che da oche ha come contenuto «anatra o oca», non «anatra» — e la misrappresentazione scompare trasformandosi in rappresentazione corretta di un tipo disgiuntivo. Per preservare la possibilità dell'errore bisogna trovare un modo per asimmetrizzare le condizioni causali: dire che in qualche senso normativo le anatre sono le *vere* cause dello stato, mentre le oche sono cause spurie, parassite, deviate.

Il problema del contenuto semantico nel suo senso più tecnico riguarda dunque due questioni distinte che è opportuno separare prima di procedere, poiché una parte significativa della confusione nella letteratura deriva dal loro intreccio non riconosciuto. La prima è la questione della *determinatezza* del contenuto: perché uno stato mentale ha il contenuto che ha, e non uno più ampio, più ristretto o radicalmente diverso? Perché il mio stato visivo quando guardo una mela è *di una mela* e non *di un frutto*, o *di una sorgente di fotoni di lunghezza d'onda compresa tra 600 e 700 nanometri*? La seconda è la questione della *normatività* del contenuto: che cosa distingue la rappresentazione corretta da quella scorretta? In virtù di che cosa uno stato può essere in errore rispetto al suo oggetto? Queste due questioni sono connesse — una risposta alla seconda presuppone una risposta alla prima — ma non sono identiche, e come si vedrà le principali teorie naturalistiche del contenuto tendono a privilegiare l'una a scapito dell'altra.

I.2 — Frege, Russell e la doppia struttura del contenuto: riferimento, senso e descrizione

Prima di affrontare le teorie naturalistiche contemporanee, è indispensabile risalire alle origini filosofico-logiche del problema del contenuto per comprendere appieno la struttura delle difficoltà che quelle teorie si trovano ad affrontare. La svolta fregeana del 1892 — con la distinzione tra Sinn e Bedeutung nell'articolo «Über

Sinn und Bedeutung» — ha stabilito i termini fondamentali entro cui il dibattito si è poi svolto, e molte delle controversie contemporanee non sono comprensibili senza una conoscenza almeno sommaria di quella svolta.

Il problema che motiva la distinzione fregeana è il seguente. Se il significato di un'espressione singolare (un nome proprio, un termine referenziale) coincidesse con il suo riferimento — con l'oggetto che denota — allora due espressioni co-referenziali dovrebbero avere lo stesso significato, e dunque essere intercambiabili in ogni contesto senza alterare il valore di verità degli enunciati in cui compaiono. Ma «l'Stella del Mattino è la Stella della Sera» — che è un'identità tra i due nomi di Venere — esprime una scoperta astronomica, mentre «l'Stella del Mattino è l'Stella del Mattino» è una tautologia priva di contenuto informativo. Se le due espressioni avessero lo stesso significato, le due proposizioni avrebbero lo stesso valore cognitivo, il che è evidentemente falso. La distinzione tra Sinn (senso, o modo di presentazione) e Bedeutung (riferimento, o denotazione) risolve questa difficoltà: due espressioni possono avere lo stesso riferimento ma sensi diversi, e i sensi diversi spiegano la diversa portata cognitiva degli enunciati.

La distinzione fregeana ha implicazioni profonde per la teoria del contenuto mentale. Se il contenuto di uno stato mentale è determinato dal senso — dal modo di presentazione dell'oggetto — piuttosto che dal solo riferimento, allora il contenuto mentale ha una struttura intensionale che non si esaurisce nella relazione con l'oggetto. Due soggetti che pensano allo stesso oggetto sotto descrizioni diverse hanno stati mentali con lo stesso riferimento ma con contenuti diversi, e questa diversità è semanticamente rilevante. La domanda che sorge naturalmente è: che cosa determina il senso, il modo di presentazione, attraverso cui un soggetto coglie il suo riferimento? È una proprietà interna della mente del soggetto, o dipende anche da fattori esterni? È questa la domanda che ha diviso internalisti ed externalisti nella filosofia della mente degli ultimi decenni.

Bertrand Russell, dal canto suo, ha elaborato una risposta radicalmente diversa al problema dei nomi propri e della referenza attraverso la teoria delle descrizioni — un contributo che ha avuto un'influenza quasi pari a quella fregeana sul successivo sviluppo della filosofia del linguaggio. Russell sosteneva che i nomi propri ordinari — «Napoleone», «Platone», «Pegaso» — non siano logicamente nomi propri genuini ma abbreviazioni di descrizioni definite: dire «Napoleone fu sconfitto a Waterloo» equivale logicamente a dire «l'individuo che fu imperatore dei Francesi e comandante a Austerlitz ... fu sconfitto a Waterloo». I nomi propri genuini — i *logically proper names* — sono per Russell termini che si riferiscono direttamente a oggetti con cui il parlante è in rapporto di conoscenza diretta (*acquaintance*): i dati sensoriali immediati e

forse il sé come soggetto di esperienza. Tutto il resto è oggetto di conoscenza per descrizione, mediata da predicati e proprietà.

La teoria russelliana ha implicazioni contrastanti per il problema del contenuto semantico. Da un lato, l'idea che la maggior parte del nostro riferimento sia mediato da descrizioni introduce un elemento di indeterminatezza: se «Napoleone» abbrevia una descrizione, il suo contenuto dipende da quale descrizione costituisce il significato del nome per il parlante dato, e diversi parlanti potrebbero associare descrizioni diverse allo stesso nome, con conseguenti problemi di intersoggettività e comunicazione. Dall'altro, la teoria russelliana offre una risposta al problema del riferimento ai non-esistenti: «Pegaso non esiste» non è una contraddizione in termini (come sarebbe se «Pegaso» si riferisse a qualcosa e di quel qualcosa si affermasse la non-esistenza) ma l'affermazione che la descrizione associata a «Pegaso» non è istanziata da nessun individuo.

La critica di Kripke a entrambe le posizioni — fregeana e russelliana — nella sua forma sistematica delle *Naming and Necessity* (1972) ha riconfigurato il campo in modo che continua a essere determinante. Kripke ha sostenuto che i nomi propri non sono descrizioni abbreviate ma *designatori rigidi*: espressioni che denotano lo stesso individuo in tutti i mondi possibili in cui quell'individuo esiste. «Napoleone» non è sinonimo di nessuna descrizione: è un'etichetta causalmente ancorata all'individuo che è stato battezzato con quel nome, e la catena causale che connette il nome all'individuo battezzato è ciò che costituisce la referenza, non l'associazione di una descrizione. Il contenuto del nome non è un modo di presentazione fregeano ma un riferimento diretto, e le proprietà che associamo al nome appartengono alla nostra conoscenza dell'individuo denotato, non al significato del nome.

Questa tripartizione di posizioni — la teoria del senso fregeana, la teoria descrittivista russelliana, il riferimento diretto kripkeano — costituisce lo sfondo rispetto al quale le teorie naturalistiche del contenuto si sono sviluppate. Come si vedrà, ognuna di queste posizioni trova echi nelle teorie contemporanee: la semantica informazionale di Dretske ha affinità con il riferimento diretto kripkeano, la teleosemantica di Millikan incorpora elementi descrittivisti attraverso la nozione di condizione di proprietà normale, e la codifica predittiva gerarchica offre una quarta alternativa che supera la tripartizione classica attraverso la struttura del modello generativo.

I.3 — La svolta computazionale e il problema della semantica interna

La rivoluzione cognitivista degli anni cinquanta e sessanta ha trasformato il problema dell'intenzionalità in un problema ingegneristico prima ancora di risolverlo come problema filosofico. L'idea centrale del cognitivismo classico — articolata con massima

nitidezza da Fodor nel suo programma del *Language of Thought* — è che la mente sia un sistema di elaborazione di simboli con contenuto semantico, e che i processi cognitivi siano computazioni definite sulla struttura sintattica di questi simboli. La sintassi — le proprietà formali dei simboli — è causalmente efficace; la semantica — il contenuto di quei simboli — superviene sulla sintassi in modo tale che i processi sintatticamente corretti sono semanticamente coerenti. La computazione è, per usare la metafora fodoriana, semantica guidata dalla sintassi.

Questa impostazione sposta ma non dissolve il problema del contenuto. Se la mente manipola simboli, i simboli hanno contenuto semantico, e la computazione è sensibile alla struttura formale dei simboli in modo tale da preservare la coerenza semantica, rimane aperta la domanda: in virtù di che cosa i simboli mentali hanno il contenuto che hanno? L'impostazione computazionale fornisce una risposta parziale: il contenuto è in parte determinato dal *ruolo inferenziale* del simbolo, dal suo posto nella rete delle relazioni computazionali con altri simboli. Ma il ruolo inferenziale da solo non è sufficiente: due simboli possono avere lo stesso ruolo inferenziale in due sistemi radicalmente diversi e tuttavia avere contenuti diversi. Il simbolo ACQUA nel sistema cognitivo di un essere umano cresciuto sulla Terra e il simbolo ACQUA nel sistema cognitivo di un essere umano cresciuto su un pianeta gemello (il *Twin Earth* di Putnam) hanno lo stesso ruolo inferenziale ma, secondo Putnam, contenuti diversi: il primo si riferisce a H_2O , il secondo a XYZ. Il contenuto, nella formulazione celebre di Putnam del 1975, non è *nella testa*: dipende dalle relazioni causali e storiche che il sistema cognitivo intrattiene con il suo ambiente. Questa è la tesi dell'esternalismo del contenuto, che ha profonde implicazioni per qualsiasi teoria computazionale della mente.

L'esperimento mentale del *Twin Earth* di Putnam, per quanto ampiamente discusso, merita una presentazione precisa perché le sue implicazioni sono spesso fraintese. Oscar e il suo gemello Oscar, *residenti rispettivamente sulla Terra e sul Twin Earth, hanno storie psicologiche identiche: stesse credenze, stessi desideri, stesso ruolo inferenziale per ogni simbolo nel loro lessico mentale. In particolare, Oscar e Oscar hanno lo stesso ruolo inferenziale per il simbolo ACQUA: credono che l'acqua sia bagnata, che si trovi nei fiumi, che cada come pioggia, che estingua la sete. Eppure, sostiene Putnam, il loro stato mentale è diverso: Oscar si riferisce a H_2O , Oscar* si riferisce a XYZ. La differenza non è nella testa ma nel mondo: l'acqua con cui Oscar è in contatto causale è H_2O , l'acqua con cui Oscar* è in contatto causale è XYZ. La morale putnamiana — che il significato non è nella testa — ha avuto un impatto enorme sulla filosofia della mente, stabilendo l'esternalismo semantico come posizione di default per molti filosofi analitici.*

Tyler Burge ha radicalizzato ulteriormente la posizione esternalista attraverso il suo esperimento mentale sul «soggetto artritico». Un

soggetto che conosce la parola «artrite» attraverso contatti linguistici con la comunità medica ma ha sviluppato una comprensione parzialmente errata — crede che l'artrite possa colpire anche i muscoli, non solo le articolazioni — ha credenze su ciò che la comunità chiama «artrite», non su ciò che lui individualmente chiamerebbe con quel nome se avesse definito il termine in modo autonomo. Il contenuto delle sue credenze è determinato dalle pratiche linguistiche della comunità, non dalla struttura interna della sua mente. Questa forma di esternalismo sociale — distinta dall'esternalismo naturalistico di Putnam — pone il problema del contenuto in una prospettiva comunitaria e normativa che le teorie puramente causali faticano a incorporare.

La tensione tra internalismo e esternalismo permea tutta la filosofia della mente degli ultimi cinquant'anni. Il cognitivismo classico tende all'internalismo: il contenuto è determinato dalla struttura funzionale interna del sistema. L'esternalismo — nelle sue varianti causale (Dretske, Stampe), teleologica (Millikan, Papineau), contestuale (Putnam, Burge) — sostiene invece che il contenuto dipende da relazioni che il sistema intrattiene con il mondo esterno o con la sua storia causale ed evolutiva. Il problema della codifica predittiva gerarchica, come si argomenterà nei capitoli centrali, offre risorse per una posizione terza che supera questa dicotomia: il contenuto è determinato dalla struttura del modello generativo che il sistema ha costruito del suo ambiente, e questo modello è allo stesso tempo interno al sistema (è codificato nelle connessioni sinaptiche e nelle distribuzioni di probabilità a priori rappresentate ai vari livelli gerarchici) ed essenzialmente relazionale rispetto all'ambiente (è costruito e affinato dall'interazione con quell'ambiente). La distinzione tra interno ed esterno, nelle reti predittive, cede il posto alla distinzione tra prior e likelihood — una distinzione tecnica che, come si mostrerà, ha implicazioni filosofiche di notevole portata.

La nozione di ruolo funzionale o ruolo inferenziale merita una trattazione più approfondita di quella che le viene solitamente riservata nelle presentazioni del problema del contenuto. La teoria del ruolo funzionale del contenuto — sviluppata nella sua forma più rigorosa da Ned Block con la nozione di funzionalismo della scatola nera e da Brian Loar e Gilbert Harman con la nozione di ruolo concettuale — sostiene che il contenuto di uno stato mentale è determinato dal suo ruolo causale-inferenziale nel sistema cognitivo: dalle sue relazioni causali con gli input sensoriali, con altri stati mentali, e con gli output comportamentali. Questa posizione ha il merito di evitare il mistero del dualismo — il contenuto non è un'entità separata dai processi neurali ma una proprietà dei processi stessi — ma genera immediatamente il problema della sistematicità olistica: se il contenuto è determinato dall'intero ruolo inferenziale, allora due sistemi con ruoli inferenziali leggermente diversi hanno contenuti diversi, e la comunicazione tra sistemi con storie di apprendimento diverse sembra richiedere una condivisione quasi perfetta dei ruoli

inferenziali che difficilmente può essere garantita. Questo problema della particolarità olistica del contenuto funzionale è uno dei motivi per cui molti filosofi hanno cercato un fondamento del contenuto esterno al sistema piuttosto che nel solo ruolo inferenziale.

I.4 — Intenzionalità originaria e intenzionalità derivata: una distinzione imprescindibile

Prima di esaminare le principali teorie naturalistiche, è indispensabile affrontare una distinzione che Searle ha reso esplicita — benché già implicita in molta della tradizione precedente — e che divide il campo in modo che nessuna teoria può ignorare: la distinzione tra intenzionalità *originaria* e intenzionalità *derivata*. L'intenzionalità derivata è quella che possiedono segni, simboli e rappresentazioni artificiali: la parola scritta «gatto» si riferisce ai gatti, ma questa referenza le è *attribuita* dagli utenti del linguaggio. Il testo scritto non ha intenzionalità propria; la sua intenzionalità gli deriva dall'intenzionalità dei parlanti che lo producono e lo interpretano. L'intenzionalità originaria, al contrario, sarebbe quella degli stati mentali stessi: il mio pensiero *di un gatto* non deriva la sua intenzionalità da nessuna fonte esterna, la possiede costitutivamente.

Searle ha usato questa distinzione per argomento contro le teorie computazionali forti della mente: i sistemi computazionali possono avere al più intenzionalità derivata, attribuita dagli osservatori umani che assegnano contenuto ai loro stati. Il famoso esperimento mentale della stanza cinese mira a mostrare che la manipolazione sintattica di simboli — per quanto sofisticata e comportamentalmente adeguata — non è sufficiente per generare intenzionalità genuina, originaria. La risposta cognitivista standard è che l'esperimento mentale è fuorviante perché considera la capacità semantica di un componente isolato (l'agente nella stanza) invece di quella del sistema complessivo. Ma questa risposta, per quanto plausibile, non risponde alla questione di fondo: che cosa rende originaria l'intenzionalità di un sistema? Che cosa distingue un sistema che ha veramente stati mentali *di* qualcosa da un sistema che simula tale relazione?

La critica di Searle, nella sua versione più sofisticata, non è semplicemente che i computer mancano di qualcosa di ineffabile. È piuttosto che la semantica — il contenuto — non è una proprietà formale ma una proprietà biologicamente radicata. Solo certi tipi di sistemi biologici — sistemi con la struttura causale appropriata, con la storia evolutiva appropriata, con il tipo di interazione con il mondo appropriato — possono avere intenzionalità originaria. Questa posizione, che Searle chiama realismo biologico, è meno mistica di quanto possa sembrare: non postula nessuna proprietà non fisica, ma afferma che la biologia — la struttura causale specifica dei sistemi neurali — è la condizione necessaria dell'intenzionalità originaria, così come la struttura

biochimica del cuore è la condizione necessaria della funzione di pompare il sangue.

La proposta sviluppata in questo lavoro non risolve questa questione nella sua forma metafisica più radicale — che probabilmente non ammette soluzione puramente computazionale — ma la riformula in termini che permettono di distinguere, all'interno della classe dei sistemi computazionali, quelli la cui struttura interna è tale da supportare una semantica naturalmente ancorata al mondo da quelli in cui la semantica rimane puramente convenzionale. Il modello generativo di una rete predittiva gerarchica non è un insieme arbitrario di regole di manipolazione sintattica: è un modello causale del mondo, costruito per minimizzare la sorpresa sotto il vincolo che le predizioni siano generate da strutture latenti reali. Questa struttura causale interna al modello — il fatto che le distribuzioni a priori e le funzioni di verosimiglianza codifichino assunzioni sulla struttura causale dell'ambiente — è la sorgente di un ancoraggio semantico che va oltre la semplice correlazione.

È opportuno distinguere, a questo punto, tre versioni della critica dell'intenzionalità derivata che circolano spesso in modo indifferenziato nella letteratura. La prima versione, la più radicale, sostiene che nessun sistema artificiale — nessun sistema computazionale costruito dall'uomo — potrebbe avere intenzionalità originaria. Questa versione dipende da assunzioni forti sul ruolo della biologia e non segue semplicemente dall'argomento della stanza cinese. La seconda versione, più moderata, sostiene che l'intenzionalità originaria richiede qualcosa che i sistemi di manipolazione sintattica puri non hanno: non necessariamente la biologia in senso stretto, ma qualcosa come una connessione causale diretta con il mondo, una storia di interazione con l'ambiente, che trasformi i processi computazionali da pura manipolazione formale in qualcosa di semanticamente ancorato. La terza versione, la più debole, sostiene semplicemente che l'intenzionalità non è un fenomeno monolitico e che le diverse teorie del contenuto catturano aspetti diversi di un fenomeno multidimensionale. Come si argomenterà, la teleosemantica computazionale sincronica abbraccia una versione raffinata della seconda posizione, distinguendo tra il processo sintattico puro — che ha al più intenzionalità derivata — e il modello generativo costruito attraverso l'interazione causale con l'ambiente — che possiede una forma di intenzionalità originaria debole nel senso che sarà precisato nel Capitolo VII.

I.5 — Quine, l'indeterminatezza della traduzione e la naturalizzazione dell'epistemologia

Prima di esaminare le principali teorie naturalistiche del contenuto, è necessario affrontare la sfida più radicale al progetto stesso di una teoria naturalistica del contenuto determinato: la tesi quiniiana dell'indeterminatezza della traduzione, che minaccia di dissolvere il

problema del contenuto piuttosto che di risolverlo, concludendo che non esiste un fatto della questione circa il contenuto delle credenze e il significato delle espressioni linguistiche.

Willard Van Orman Quine ha sostenuto, in *Word and Object* (1960) e in numerosi lavori successivi, che il processo di traduzione radicale — il tentativo di interpretare il linguaggio di una comunità sconosciuta a partire solo dall'evidenza del comportamento linguistico in contesti osservabili — non è mai in grado di determinare univocamente il manuale di traduzione corretto. Due manuali di traduzione empiricamente equivalenti — in grado di rendere conto con uguale precisione del comportamento osservabile dei parlanti — potrebbero tuttavia attribuire significati radicalmente diversi alle espressioni della lingua. Il celebre esempio del «gavagai»: uno studioso di campo sul campo con una tribù sconosciuta osserva un parlante pronunciare «gavagai» al passaggio di un coniglio. La traduzione naturale è «coniglio», ma potrebbero essere ugualmente compatibili con l'evidenza osservabile le traduzioni «parte non separata di coniglio», «stadio temporale di coniglio», «conigliosità impersale», e così via. Nessuna quantità di evidenza comportamentale può determinare quale delle traduzioni sia quella corretta, perché l'evidenza è compatibile con tutte. La stessa indeterminatezza vale — e questa è la tesi più radicale — per il nostro stesso linguaggio: non esiste un fatto della questione circa ciò che le nostre espressioni significano, oltre a ciò che è determinato dalla nostra propensione all'accordo verbale.

La sfida quiniana è particolarmente acuta per le teorie naturalistiche del contenuto che vogliono fondare la determinatezza del contenuto su relazioni causali o informazionali con il mondo. Se il comportamento linguistico osservabile non è sufficiente a determinare il manuale di traduzione, e se è solo il comportamento linguistico che è accessibile all'interpretazione dall'esterno, allora le relazioni causali con il mondo non determinano il contenuto in modo univoco: le stesse relazioni causali sono compatibili con manuali di traduzione diversi, ovvero con attribuzioni di contenuto diverse. Il naturalismo del contenuto, che vuole determinare il contenuto attraverso relazioni naturali tra sistema e mondo, non sfugge all'indeterminatezza quiniana perché la struttura causale del mondo è compatibile con molteplici schemi di interpretazione.

La risposta che le teorie naturalistiche del contenuto hanno elaborato alla sfida quiniana prende due forme principali. La prima, adottata da Millikan, è di abbracciare un realismo robusto sulle funzioni proprie che fornisce una base per la determinatezza del contenuto indipendente dall'evidenza comportamentale: il contenuto di una rappresentazione è determinato dalla storia selettiva del sistema, e questa storia è un fatto empirico della biologia evolutiva, non una proiezione interpretativa dell'osservatore. La seconda, adottata da vari autori neodretsiani, è di sostenere che le relazioni informazionali sono abbastanza

precise da vincolare il manuale di traduzione in modo sufficiente, anche se non in modo assoluto: la struttura delle dipendenze probabilistiche tra stati del sistema e stati del mondo non è compatibile con qualsiasi schema interpretativo, ma solo con un numero limitato. La teleosemantica computazionale sincronica, come si vedrà, offre una terza risposta fondata sulla struttura del modello generativo: il modello determina il contenuto attraverso la struttura delle predizioni che genera, e questa struttura è sufficientemente specifica da vincolare l'interpretazione in modo che va al di là della semplice osservazione del comportamento.

I.6 — Il problema della normatività e l'insufficienza del puro naturalismo causale

Prima di passare all'esame delle teorie specifiche, vale la pena di fermarsi sulla questione della normatività, poiché essa costituisce il vincolo più stringente che qualsiasi teoria del contenuto semantico deve soddisfare. La normatività del contenuto consiste nel fatto che gli stati rappresentazionali possono essere *corretti* o *scorretti* rispetto ai loro oggetti: una percezione può essere illusoria, una credenza può essere falsa, un'aspettativa può essere delusa. La normatività non è una caratteristica di tutti i processi causali: un termostato che registra una temperatura di 20°C quando la temperatura è effettivamente di 20°C non è più *corretto* di quanto sarebbe *scorretto* se registrasse 25°C — è semplicemente in uno stato causato da certe condizioni fisiche. Per parlare di correttezza e scorrettezza bisogna che esista uno standard normativo rispetto al quale valutare lo stato del sistema. Dove si trova questo standard normativo in un sistema fisico naturale?

Il problema della normatività ha una struttura che Wittgenstein aveva anticipato nelle *Philosophical Investigations* attraverso la discussione delle regole e del seguire una regola. Wittgenstein osservava che nessuna interpretazione di una regola può, da sola, determinare quale applicazione della regola sia corretta: qualsiasi comportamento può essere interpretato come conforme a qualche regola, e qualsiasi regola può essere interpretata in modo tale da giustificare qualsiasi comportamento. Il paradosso del seguire una regola — che Kripke ha formalizzato e reso celebre nella sua forma dello scettico — è che non esiste un fatto della questione circa quale regola il parlante stia seguendo che sia indipendente dalla pratica sociale in cui il seguire regole è incorporato. Questo paradosso wittgensteiniano ha risonanze con il problema della normatività del contenuto: in entrambi i casi, la normatività non può essere localizzata in nessun fatto naturale isolato ma richiede un contesto di pratiche sociali e normative che la trascende.

Le risposte principali che la tradizione analitica ha elaborato si articolano attorno a tre candidati: la struttura causale contraffattuale (Dretske: lo stato è corretto quando è causato da ciò che *dovrebbe* causarlo in condizioni normali), la storia selettiva (Millikan: lo stato è corretto quando svolge la funzione per cui è

stato selezionato), e la struttura del desiderio o della preferenza (varianti teleologiche diverse). Ciascuna di queste risposte cattura qualcosa di genuino ma enfrenta difficoltà specifiche che si esamineranno nei capitoli successivi. Quello che importa rilevare in questo contesto introduttivo è la struttura generale del problema: qualsiasi teoria naturalistica del contenuto deve rispondere alla domanda *in che senso normativo* uno stato rappresenta ciò che rappresenta, e questa risposta deve essere formulata senza circolarità e senza postulare entità non naturali.

La tesi che si difenderà è che il framework della codifica predittiva offre, per la prima volta in modo tecnicamente esplicito, una risposta alla questione della normatività che non dipende né dalla storia selettiva né da relazioni causali contraffattuali astratte, ma dalla struttura interna del modello generativo stesso: uno stato rappresentativo è corretto quando il modello generativo che lo ha prodotto assegna alta verosimiglianza alle osservazioni date lo stato in questione, e scorretto quando l'errore di predizione — la discrepanza tra le osservazioni effettive e quelle predette dal modello — è elevato. La normatività è endogena al processo computazionale, non importata da una storia evolutiva esterna o da relazioni causali ideali. Questa internalizzazione della normatività è il contributo più originale della prospettiva predittiva al problema del contenuto semantico, e la sua elaborazione rigorosa costituisce il cuore della proposta qui sviluppata.

Vale la pena sottolineare che la normatività endogena non è identica alla mera coerenza interna. Un sistema potrebbe essere internamente coerente — tutte le sue rappresentazioni si accordano tra loro in modo tale da minimizzare la mutua incoerenza — pur essendo globalmente disconnesso dal mondo. La normatività del processo predittivo è diversa: è la normatività di un sistema che è in contatto causale con il mondo attraverso i suoi segnali sensoriali, e che valuta la correttezza delle sue rappresentazioni rispetto a quei segnali. La sorpresa sensoriale — il segnale di errore che indica quanto le previsioni del modello si discostano dalle osservazioni effettive — è ancorata al mondo fisico, non solo alla struttura interna del sistema. Questa struttura bifronte del processo predittivo — ancorata al mondo attraverso i segnali sensoriali, ma normativa rispetto alla struttura del modello — è ciò che permette di distinguere la normatività predittiva dalla mera coerenza interna.

I.7 — Davidson, la causazione mentale e l'anomalismo del mentale

Un nodo concettuale che attraversa tutta la tradizione della filosofia della mente analitica, e che ha implicazioni dirette per il problema del contenuto semantico, è il problema della causazione mentale: in che senso gli stati mentali — credenze, desideri, intenzioni — possono esercitare efficacia causale in un mondo fisicamente chiuso? Il problema è stato articolato con massima

chiarezza da Davidson nella sua teoria del monismo anomalo (1970) e da Kim nella sua critica al dualismo di proprietà (1989-2005).

Davidson sosteneva che gli stati mentali siano stati fisici descritti sotto tipi mentali, e che le relazioni causali tra stati mentali siano in realtà relazioni causali tra gli stati fisici corrispondenti. Il mentale è anomalo — non si sottomette a leggi strettamente nomologiche, a differenza del fisico — ma è causalmente efficace nella misura in cui ogni evento mentale è identico a un evento fisico che entra in relazioni causali regolate da leggi fisiche. Questa posizione, per quanto elegante, genera la preoccupazione che il mentale sia causalmente epifenomenico in quanto *mentale*: le relazioni causali sono relazioni tra eventi fisici descritti in termini fisici, e la descrizione mentale degli stessi eventi è causalmente irrilevante. La distinzione tra la ragione di un'azione — le credenze e i desideri che la razionalizzano — e la sua causa — lo stato fisico che la produce — diventa problematica: la ragione è causalmente efficace solo nella misura in cui coincide con la causa fisica, non in virtù del suo contenuto.

Kim ha radicalizzato questa preoccupazione attraverso l'argomento dell'esclusione: se ogni effetto fisico ha una causa fisica sufficiente (chiusura causale del fisico), e se le proprietà mentali sono distinte dalle proprietà fisiche (non-riduzionismo), allora le proprietà mentali non possono contribuire causalmente agli effetti fisici senza violare la chiusura causale o produrre una overdeterminazione sistematica della causalità. Il mentale sembra destinato all'epifenomenismo: le proprietà mentali accompagnano i processi fisici senza contribuire causalmente a nessuno di essi.

Il problema della causazione mentale ha una connessione diretta con il problema del contenuto semantico attraverso la nozione di causalità semantica o causazione mediante contenuto. Se le credenze causano le azioni attraverso il loro contenuto — non semplicemente attraverso le loro proprietà fisiche ma in virtù di *cio che rappresentano* — allora la causalità mentale richiede che il contenuto stesso sia causalmente rilevante, non solo il substrato fisico che lo realizza. Ma se il contenuto è determinato da relazioni storiche o sociali che il substrato fisico non cattura completamente, come può il contenuto — e non solo il substrato — essere causalmente efficace? Questa domanda — in che senso il contenuto semantico ha efficacia causale — è una delle questioni più profonde della filosofia della mente e non ammette una risposta semplice.

La soluzione che il framework della codifica predittiva suggerisce, e che sarà elaborata nel Capitolo IX, è di tipo strutturalistico: il contenuto è causalmente efficace nella misura in cui è codificato nella struttura del modello generativo, e il modello generativo — con le sue funzioni generative, le sue distribuzioni a priori, le sue matrici di precisione — è parte del sistema fisico che produce il comportamento. Non si tratta di una riduzione del contenuto al

substrato fisico, ma di una ricollocazione del contenuto nella struttura del processo computazionale che il substrato fisico implementa. Il contenuto non fluttua sopra il substrato come un'entità separata: è incorporato nella struttura funzionale del modello generativo, che è a sua volta implementata nel substrato neurale attraverso i pesi sinaptici e i pattern di connettività. Questa identificazione tra struttura generativa e substrato fisico non è un'identificazione tipo-tipo nel senso del riduzionismo classico: è un'identificazione strutturale che lascia aperta la possibilità di realizzazioni multiple e di variazione individuale nella struttura del modello.

II — La teleosemantica di Millikan: funzioni proprie, storia selettiva e i limiti del fondamento biologico

II.1 — L'intuizione di base: il contenuto come funzione biologica

Tra le teorie naturalistiche del contenuto semantico elaborate nella tradizione analitica, quella di Ruth Garrett Millikan occupa una posizione di speciale rilievo per la profondità dell'impostazione, la coerenza interna e la capacità di affrontare direttamente il problema della misrappresentazione che vanifica le teorie puramente causali. Millikan ha sviluppato il proprio programma in *Language, Thought, and Other Biological Categories* (1984) e in una serie di elaborazioni successive, tra cui *White Queen Psychology and Other Essays for Alice* (1993) e *On Clear and Confused Ideas* (2000), fondando la semantica dei sistemi rappresentazionali sulla nozione biologica di *funzione propria* — una nozione tecnica che Millikan distingue con cura dalle accezioni ordinarie di funzione e che costituisce il cardine dell'intera costruzione.

L'intuizione di partenza è che la correttezza e la scorrettezza dei rappresentazioni mentali siano fenomeni dello stesso tipo della correttezza e scorrettezza nel funzionamento degli organi biologici. Il cuore ha la funzione di pompare il sangue: questa funzione non è assegnata dall'osservatore, ma è una proprietà obiettiva dell'organo determinata dalla sua storia selettiva. Un cuore che non pompa sangue è difettoso, malfunzionante — e questo giudizio normativo è fondato nella biologia evolutiva, non nella convenzione. Il cuore esiste, nella sua forma attuale, perché i cuori ancestrali pompavano il sangue e questo pompaggio contribuiva alla sopravvivenza e alla riproduzione degli organismi che li possedevano; la selezione naturale ha favorito le varianti che assolvevano questa funzione meglio di altre. Questa storia selettiva è ciò che conferisce alla funzione il suo carattere normativo: il cuore *deve* pompare sangue non in senso logico o matematico, ma

nel senso che è questa la prestazione per cui esiste, la prestazione il cui svolgimento spiega la sua presenza nell'organismo.

La posizione di Millikan si distingue nettamente dalla posizione funzionalista nella filosofia della mente che identifica le proprietà mentali con i ruoli causali o funzionali che istanziano. Il funzionalismo è una posizione sincronicamente definita: lo stato mentale *S* è definito dai suoi rapporti causali attuali con gli input, con altri stati, e con gli output. Millikan, al contrario, adotta una prospettiva diacronicamente definita: la funzione propria di un dispositivo non è determinata dal suo ruolo causale attuale ma dalla sua storia di riproduzione e selezione. Un cuore che non pompa sangue a causa di un'ostruzione ha ancora la funzione propria di pompare sangue: la funzione non è eliminata dall'incapacità attuale di svolgerla, poiché la funzione è determinata dalla storia, non dallo stato attuale. Questa distinzione è cruciale per il problema della misrappresentazione: una rappresentazione può avere la funzione propria di tracciare una certa condizione del mondo anche quando non riesce a farlo, esattamente come un cuore difettoso ha ancora la funzione di pompare sangue anche quando non ci riesce.

Millikan estende questa struttura alle rappresentazioni mentali e agli atti linguistici. I sistemi rappresentazionali — credenze, percezioni, desideri, atti linguistici — hanno funzioni proprie determinate dalla loro storia selettiva. La storia qui può essere evolutiva in senso stretto (filogenetica) o può essere una storia di apprendimento individuale modellata da meccanismi selettivi (rinforzo, abitudine, correzione dell'errore): Millikan usa il termine *storia di riproduzione propria* per designare qualsiasi storia in cui un meccanismo è stato selezionato o rinforzato in virtù di certe prestazioni. Una credenza vera è una credenza che assolve la funzione propria del sistema dossastico — la funzione di guidare il comportamento in modo efficace rispetto all'ambiente. Una credenza falsa è una credenza che non assolve questa funzione, che è difettosa nello stesso senso in cui è difettoso un cuore che non pompa sangue.

II.2 — La struttura formale delle funzioni proprie

La nozione di funzione propria richiede una formulazione precisa, poiché la sua vaghezza intuitiva ha spesso portato a fraintendimenti nelle discussioni sulla teleosemantica. Millikan distingue tra funzioni proprie *dirette* e funzioni proprie *derivate*. Una funzione propria diretta è quella che il dispositivo ha perché i suoi predecessori hanno svolto quella funzione e questo svolgimento ha contribuito alla riproduzione dei predecessori stessi. L'occhio ha la funzione propria diretta di vedere perché gli occhi ancestrali vedevano e la visione contribuiva alla sopravvivenza. Una funzione propria derivata è quella che un dispositivo ha in virtù di un sistema di relazioni più complesso: le parti di un organismo hanno funzioni proprie derivate che contribuiscono alla funzione propria

diretta dell'organismo come un tutto; i segni linguistici hanno funzioni proprie derivate che contribuiscono alle funzioni comunicative del linguaggio.

La definizione ricorsiva di funzione propria che Millikan elabora può essere sintetizzata nel modo seguente. Sia D un dispositivo appartenente a una famiglia di dispositivi riprodotti (nel senso ampio che include sia la riproduzione biologica sia la produzione secondo un modello). D ha la funzione propria di produrre effetto E se e solo se D esiste come membro della sua famiglia almeno in parte perché i dispositivi di quella famiglia hanno in passato prodotto E, e la produzione di E ha contribuito alla loro riproduzione o alla riproduzione del sistema di cui fanno parte. In simboli, indicando con $F(D)$ l'insieme delle funzioni proprie di D, con $H(D)$ la storia di riproduzione di D e con $C(E, H(D))$ la condizione che E abbia contribuito causalmente a $H(D)$:

$$F(D) = \{ E : C(E, H(D)) \}$$

La dipendenza dall'intera storia di riproduzione, e non solo dall'istanza attuale o da un insieme di istanze recenti, è ciò che conferisce alle funzioni proprie il loro carattere normativo irriducibile: la norma non è stabilita dall'istanza attuale ma dall'intera distribuzione storica delle prestazioni del tipo di dispositivo.

Millikan applica questa struttura ai dispositivi rappresentazionali attraverso la nozione di *funzione propria normale*: la condizione in cui un dispositivo assolve la propria funzione propria è la condizione che Millikan chiama *condizione di proprietà normale* (Normal Condition), e che è definita come la condizione il cui verificarsi era sistematicamente associato al successo degli antenati del dispositivo nell'assolvere la funzione propria. Per i sistemi rappresentazionali, la condizione di proprietà normale è quella in cui il contenuto della rappresentazione corrisponde allo stato del mondo che la rappresentazione deve tracciare: la percezione visiva di un albero è in condizione di proprietà normale quando c'è effettivamente un albero nella posizione indicata dalla percezione.

Il quadro teorico di Millikan include anche la nozione di dispositivo *produttore* e dispositivo *consumatore*. Il sistema percettivo visivo è un produttore di rappresentazioni: genera stati rappresentazionali in risposta agli input sensoriali. Il sistema motorio e il sistema di pianificazione dell'azione sono consumatori: ricevono quelle rappresentazioni e le usano per guidare il comportamento. La funzione propria delle rappresentazioni visive è determinata non solo dalle relazioni con il sistema produttore ma anche dalle relazioni con il sistema consumatore: una rappresentazione visiva assolve la propria funzione propria quando il sistema consumatore — agendo sulla base di essa — produce comportamenti appropriati alle condizioni che la rappresentazione traccia. Questa struttura produttore-consumatore permette a Millikan di spiegare la

sistematicità del contenuto: una rappresentazione ha il contenuto che ha non perché sia causalmente correlata con una proprietà del mondo, ma perché il suo produttore è stato selezionato per generarla in risposta a quella proprietà e il suo consumatore è stato selezionato per agire su di essa come se quella proprietà valesse.

II.3 — Il contenuto come condizione di soddisfazione in condizioni normali

La nozione di contenuto semantico nella teleosemantica di Millikan è strettamente intrecciata a quella di funzione propria. Il contenuto di una rappresentazione R non è determinato da ciò che causa R nell'istanza attuale, né da ciò con cui R è correlata in modo statisticamente prevalente: è determinato da ciò che R *deve* tracciare affinché assolvere la sua funzione propria in condizioni normali. La condizione che R deve tracciare per assolvere la propria funzione propria è la *condizione di correttezza* di R , e il contenuto di R è precisamente quella condizione.

Consideriamo un esempio paradigmatico che Millikan elabora a lungo: la danza delle api. Le api esploratrici comunicano la posizione delle sorgenti di nutrimento alle api operaie attraverso una danza il cui angolo rispetto alla verticale codifica la direzione della sorgente rispetto al sole, e la cui durata codifica la distanza. Qual è il contenuto di questa danza? Non semplicemente ciò che la causa — le condizioni di luce solare, le azioni dell'ape danzante, le interazioni sociali. Il contenuto è la posizione della sorgente di nutrimento, perché è questa la condizione che la danza deve tracciare affinché assolvere la propria funzione propria: guidare le api operaie verso il nutrimento. Se la sorgente non si trova dove la danza indica — per esempio perché l'ape danzante ha fatto un errore nel calcolo della direzione — la danza è scorretta, misrappresenta, ed è esattamente in questo senso che la teleosemantica fornisce una risposta al problema della misrappresentazione. La danza è scorretta non perché sia causalmente anomala (la sua catena causale è perfettamente regolare) ma perché non si trova nella condizione di proprietà normale: le sue cause non corrispondono alla condizione che essa deve tracciare.

Formalizziamo questa struttura in modo più esplicito. Sia R una rappresentazione, $FP(R)$ la sua funzione propria, e $CN(R)$ la condizione di proprietà normale per R . Il contenuto semantico di R — che indicheremo con $Cnt(R)$ — è la proposizione che descrive $CN(R)$ dal punto di vista della condizione mondo-dipendente che deve essere soddisfatta affinché R assolvere $FP(R)$:

$$Cnt(R) = \{ w \in W : CN(R) \text{ è soddisfatta in } w \}$$

dove W è lo spazio dei mondi possibili rilevanti e la condizione di correttezza è l'insieme dei mondi in cui la condizione di proprietà

normale è soddisfatta. R è corretta nel mondo attuale w^* se e solo se $w^* \in \text{Cnt}(R)$, e scorretta se $w^* \notin \text{Cnt}(R)$.

Questa formulazione mostra come la teleosemantica di Millikan risolva in linea di principio il problema della misrappresentazione: l'errore rappresentativo è definito come l'assenza della condizione di proprietà normale nell'istanza attuale, il che non richiede nessuna anomalia causale nell'istanza. Il dispositivo può funzionare causalmente in modo perfettamente regolare e tuttavia essere in una condizione scorretta rispetto alla sua funzione propria, esattamente come un cuore può battere regolarmente e tuttavia non pompare sangue efficacemente a causa di un'ostruzione valvolare.

La struttura produttore-consumatore aggiunge un ulteriore strato di complessità alla teoria. Il contenuto di una rappresentazione è determinato non solo dalla condizione che il produttore deve tracciare, ma anche dalla condizione che il consumatore deve trovare perché l'intera catena rappresentativa assolva la propria funzione. In una catena rappresentativa semplice, queste due condizioni coincidono: il produttore deve tracciare esattamente ciò che il consumatore deve trovare. Ma in catene più complesse — come quelle che caratterizzano la cognizione umana adulta — le condizioni possono disaccoppiarsi: il produttore genera una rappresentazione in un contesto e il consumatore la interpreta in un contesto diverso, e la funzione propria dell'intera catena è determinata dall'integrazione di entrambi i contributi.

II.4 — Neander, Papineau e le varianti della teleosemantica

La teleosemantica di Millikan non è l'unica né la prima versione della posizione teleologica nel dibattito sul contenuto semantico. Karen Neander e David Papineau hanno sviluppato varianti importanti che meritano una discussione separata, poiché illuminano aspetti diversi del programma teleologico e le sue difficoltà specifiche.

Karen Neander ha articolato la propria versione della teleosemantica concentrandosi sulla questione della selezione naturale e del suo ruolo nella determinazione delle funzioni proprie. La proposta di Neander — sviluppata in una serie di articoli a partire dagli anni ottanta — enfatizza il ruolo della selezione naturale come meccanismo che conferisce alle funzioni proprie il loro carattere normativo. La selezione naturale non è solo una storia causale che spiega la presenza di un tratto: è un processo ottimizzante che seleziona i tratti in virtù delle loro prestazioni, producendo un tipo di normatività che non è semplicemente l'eco di una storia passata ma è una normatività genuinamente prospettica — i tratti selezionati sono quelli che tendono a produrre le prestazioni appropriate in condizioni normali future. Questa enfasi sulla dimensione ottimizzante della

selezione distingue la posizione di Neander da quella di Millikan: per Neander, la funzione propria non è semplicemente ciò che il dispositivo ha *fatto* in passato, ma ciò che il dispositivo è *progettato* per fare nel senso evolutivo del termine.

David Papineau ha sviluppato una variante della teleosemantica che enfatizza il ruolo dei *desideri* e delle *credenze* nella struttura rappresentazionale. La proposta di Papineau in *Philosophical Naturalism* (1993) parte dall'osservazione che la normatività del contenuto è strettamente legata alla struttura del comportamento orientato agli scopi. Le credenze e i desideri sono rappresentazioni che interagiscono per produrre comportamento: le credenze rappresentano come stanno le cose nel mondo, i desideri rappresentano come si vorrebbe che stessero. Il comportamento è orientato agli scopi nella misura in cui è causato dalla combinazione di credenze e desideri in modo tale che — se le credenze sono corrette — il comportamento soddisfa il desiderio. Questa struttura interazionale definisce il contenuto delle credenze: una credenza ha come contenuto una condizione del mondo se e solo se, quando quella condizione vale, la credenza tende a produrre — in combinazione con i desideri appropriati — comportamenti che soddisfano quei desideri. L'account papineauiano è teleologico ma in un senso più strettamente legato alla struttura dell'azione intenzionale piuttosto che alla biologia evolutiva.

La differenza tra Millikan e Papineau riflette una tensione di fondo nella teleosemantica tra due modi di intendere la normatività: come proprietà delle strutture biologiche (Millikan) o come proprietà della struttura razionale dell'agire (Papineau). Nel primo caso, la normatività è obiettiva e indipendente dall'agente — è la normatività della selezione naturale, che non dipende da nessuna prospettiva soggettiva. Nel secondo caso, la normatività è quasi-soggettiva — dipende dalla struttura dei desideri dell'agente, anche se questi desideri possono essere descritti in termini naturalistici. Come si vedrà, la teleosemantica computazionale sincronica propone una terza forma di normatività — endogena al processo computazionale — che non coincide né con la normatività biologica di Millikan né con la normatività razionale di Papineau, pur avendo affinità con entrambe.

II.5 — Il problema della storia prossimale e la distalità del fondamento teleologico

Per quanto la teleosemantica di Millikan costituisca il contributo più sistematico alla questione del contenuto semantico nel quadro naturalistico, essa affronta una serie di difficoltà che ne limitano l'applicabilità e che, come si argomenterà, la rendono insufficiente a fondare una teoria del contenuto nelle reti predittive gerarchiche. La difficoltà più profonda — che si potrebbe chiamare il problema della storia prossimale — riguarda il rapporto tra il fondamento

distale del contenuto (la storia selettiva) e la determinazione sincronica del contenuto in ogni istanza rappresentativa.

La storia selettiva che fonda le funzioni proprie è, per definizione, una storia passata: riguarda ciò che i predecessori del dispositivo hanno fatto e perché questo ha contribuito alla loro sopravvivenza o riproduzione. Ma le rappresentazioni sono fenomeni presenti: si producono in tempo reale, in risposta a stimoli attuali, nel contesto di un sistema cognitivo che deve agire qui e ora. La domanda è: in che modo la storia passata di selezione determina il contenuto della rappresentazione attuale? La risposta di Millikan è che la storia passata ha plasmato la struttura attuale del sistema rappresentazionale — i meccanismi percettivi, i sistemi di credenza, le disposizioni dossastiche — in modo tale che questi meccanismi tendono a produrre rappresentazioni corrette nelle condizioni che erano normali nell'ambiente evolutivo. Ma questo fondamento distale genera un disallineamento sistematico tra ciò che il sistema rappresenta (determinato dalla storia selettiva) e ciò che il sistema sta effettivamente elaborando (determinato dalle condizioni ambientali attuali).

Il problema si manifesta con chiarezza nel caso degli ambienti evolutivi disadattativi. Un sistema cognitivo umano che si trovi nell'ambiente artificiale di una città moderna è esposto a stimoli che non erano presenti nell'ambiente evolutivo di adattamento. Le funzioni proprie dei suoi meccanismi percettivi sono state selezionate per tracciare le caratteristiche rilevanti dell'ambiente pleistocenico, non quelle dell'ambiente urbano. Il contenuto delle rappresentazioni visive di uno schermo televisivo — in termini di funzioni proprie — sarebbe allora o indeterminato (la televisione non era nell'ambiente di selezione) o scorretto (le rappresentazioni tracciano caratteristiche che non corrispondono alle condizioni di proprietà normale). Millikan ha risposto a questa difficoltà distinguendo tra il contenuto determinato dalla storia filogenetica e quello determinato dalla storia ontogenetica di apprendimento, suggerendo che i meccanismi di apprendimento hanno anch'essi funzioni proprie che permettono di acquisire nuovi contenuti in risposta a nuovi ambienti. Ma questa risposta complica la struttura teorica senza risolvere il problema: i meccanismi di apprendimento devono anch'essi essere fondati su funzioni proprie, generando una regressione che deve arrestarsi da qualche parte nella storia evolutiva.

Un ulteriore problema riguarda la granularità della selezione. La selezione naturale opera su variazioni fenotipiche nell'intera popolazione, non su rappresentazioni individuali nell'arco della vita di un organismo. Le pressioni selettive che hanno plasmato il sistema visivo umano hanno operato su tempi evolutivi di migliaia o milioni di generazioni, producendo meccanismi adattati alle statistiche dell'ambiente pleistocenico in modo statisticamente medio. Ma le rappresentazioni che il sistema produce in risposta a singoli stimoli non sono determinate da quelle statistiche medie:

sono prodotte da meccanismi specifici in risposta a stimoli specifici, e la correttezza di ciascuna rappresentazione non può essere valutata direttamente rispetto alle statistiche evolutive. La normatività teleologica è intrinsecamente statistica e populazionale, mentre la normatività che vogliamo attribuire alle rappresentazioni individuali è intrinsecamente token-specifica: vogliamo dire che questa rappresentazione, prodotta adesso in risposta a questo stimolo, è corretta o scorretta rispetto a questo stato del mondo. La teleosemantica fornisce il fondamento per la normatività del tipo di rappresentazione ma non per la normatività del token, e questa lacuna è sistematica.

II.6 — Il problema dell'indeterminatezza del contenuto

Un secondo ordine di difficoltà riguarda l'indeterminatezza del contenuto che sembra generare la teleosemantica. Si tratta di una difficoltà che Fodor ha articolato in modo particolarmente tagliente, e che Millikan stessa ha parzialmente riconosciuto pur proponendo risposte che non sembrano del tutto convincenti. La difficoltà può essere esposta come segue.

Sia *R* una rappresentazione prodotta dal sistema visivo umano in risposta a oggetti di una certa forma. In termini di storia selettiva, questa rappresentazione è stata selezionata perché tracciava le caratteristiche degli animali predatori (diciamo), il cui riconoscimento tempestivo era adattivamente vantaggioso. Ma gli animali predatori sono caratterizzati da una pluralità di proprietà: il profilo, il moto, la colorazione, l'odore, i suoni che producono. Quale di queste proprietà costituisce il contenuto della rappresentazione? In termini formali: tra tutte le condizioni che erano sistematicamente associate al successo adattativo nell'ambiente di selezione, quale individua la condizione di proprietà normale e dunque il contenuto?

Millikan distingue tra il *contenuto completo* e il *contenuto astratto* di una rappresentazione, e insiste che il contenuto è determinato dalla condizione la cui covariazione con la rappresentazione è stata *rilevante* per la selezione. Ma la nozione di rilevanza qui impiegata è essa stessa teleologica — rilevante per quale scopo, rispetto a quale funzione? — e rischia di introdurre una circolarità. In un ambiente dato, molte proprietà covariano sistematicamente: le zebre hanno strisce, quattro zampe, erbivoria, un certo schema locomotore. Quale di queste proprietà costituisce il contenuto della rappresentazione che le zebre producono? La risposta che il contenuto è la proprietà la cui covariazione con la rappresentazione è stata causalmente rilevante per il successo della funzione propria sembra corretta ma è difficile da applicare in modo non circolare senza un criterio indipendente di rilevanza causale.

La difficoltà si approfondisce quando si considera la struttura gerarchica dei sistemi percettivi. Un sistema visivo elabora le informazioni visive a molti livelli di astrazione: dai bordi e le frequenze spaziali (nei livelli corticali primari) agli oggetti e le categorie (nei livelli corticali associativi). A quale livello deve essere localizzato il contenuto della rappresentazione? La risposta millikaniana che il contenuto è determinato dalla funzione propria del consumatore non è del tutto soddisfacente perché la gerarchia dei consumatori è anch'essa stratificata: il sistema visivo produce output che sono consumati da molteplici sistemi a valle — il sistema dell'azione, il sistema del linguaggio, il sistema della memoria — e ciascuno di questi sistemi consumatori potrebbe essere ritenuto il consumatore «rilevante» per la determinazione del contenuto. La teleosemantica non fornisce un principio per risolvere questa ambiguità gerarchica.

Questo problema dell'indeterminatezza del contenuto è particolarmente acuto per le rappresentazioni di alto livello — credenze astratte, concetti categoriali, rappresentazioni proposizionali — la cui connessione con la storia selettiva è mediata da catene causali e funzionali lunghissime e difficilmente tracciabili. Il contenuto della credenza che la somma degli angoli di un triangolo euclideo è 180 gradi non è ovviamente determinato dalla storia selettiva in modo diretto. Millikan ha risposto estendendo la nozione di funzione propria derivata e di produttore e consumatore di segni in modo da coprire anche le rappresentazioni più astratte, ma l'estensione richiede assunzioni aggiuntive che allontanano la teoria dal suo fondamento biologico iniziale.

II.7 — Il problema dell'istantaneità: perché la storia non basta

La difficoltà più rilevante ai fini della presente indagine — che chiameremo il problema dell'istantaneità — non riguarda la coerenza interna della teleosemantica di Millikan ma la sua adeguatezza come fondamento per il contenuto nei sistemi di codifica predittiva gerarchica. La difficoltà è questa: i sistemi predittivi costruiscono e aggiornano continuamente i loro modelli generativi in risposta all'input sensoriale, e le rappresentazioni che producono possono avere contenuti determinati anche in assenza di una storia selettiva pertinente. Un sistema di codifica predittiva che per la prima volta elabora un tipo di stimolo mai incontrato prima — e questo accade costantemente nei sistemi biologici sviluppati, che si trovano in ambienti sempre parzialmente nuovi — deve poter determinare il contenuto delle sue rappresentazioni sincronicamente, cioè in base alla struttura attuale del suo modello generativo, senza fare appello a una storia selettiva che non può esistere per stimoli mai incontrati.

Più precisamente: l'aggiornamento bayesiano che caratterizza i sistemi predittivi modifica le distribuzioni di probabilità a priori in

modo tale che le rappresentazioni risultanti abbiano un contenuto determinato rispetto al modello generativo aggiornato. Se vedo per la prima volta un platipus, la mia rappresentazione visiva ha un contenuto determinato — è *di quel platipus* — anche se la mia storia selettiva non include nessuna pressione selettiva legata ai platipus. Il contenuto è determinato non dalla storia selettiva ma dall'aggiornamento del modello generativo: la mia rappresentazione mette in relazione il percelto con le categorie e le strutture inferenziali già presenti nel mio modello, producendo un contenuto che è funzione dell'intersezione tra il nuovo stimolo e la struttura preesistente del modello.

Questo argomento non mira a confutare la teleosemantica di Millikan come teoria del contenuto *in generale*, ma a mostrare che essa è insufficiente come teoria del contenuto nei sistemi predittivi. La storia selettiva può spiegare perché il sistema ha la struttura generativa che ha — perché le sue distribuzioni a priori sono quelle che sono, perché le sue funzioni di verosimiglianza hanno la forma che hanno. Ma non può spiegare il contenuto delle rappresentazioni *generate* da quella struttura in risposta a stimoli nuovi. Il contenuto in questo senso è *sincronicamente determinato* dalla struttura del modello generativo, e questa determinazione sincronica richiede una teoria che la teleosemantica, essenzialmente diacronica nel suo fondamento, non è in grado di fornire.

La distinzione tra la spiegazione della struttura del modello (diacronica, biologica) e la determinazione del contenuto a partire dal modello (sincronica, computazionale) è analoga alla distinzione tra la spiegazione dell'architettura di un sistema di calcolo (che può essere spiegata dalla storia della progettazione, dalle scelte dei costruttori, dai materiali disponibili) e la determinazione dell'output di quel sistema per un dato input (che è determinata dalla struttura del programma in esecuzione, non dalla storia del sistema). La teleosemantica spiega l'architettura; la teleosemantica computazionale sincronica spiega l'output.

II.8 — La teleosemantica come condizione necessaria ma non sufficiente

La conclusione che emerge dall'esame critico della teleosemantica di Millikan non è una confutazione globale ma una ridefinizione del suo ruolo. La storia selettiva, nella misura in cui ha plasmato la struttura del modello generativo dei sistemi cognitivi biologici, contribuisce alla spiegazione di *perché il sistema ha il tipo di capacità rappresentazionale che ha*: perché è in grado di rappresentare superfici, oggetti, agenti, relazioni causali, stati epistemic altrui. La struttura del modello generativo — la gerarchia delle distribuzioni a priori, la forma delle funzioni di verosimiglianza, i pattern di connettività che implementano l'inferenza — porta l'impronta della storia selettiva in modo che la teleosemantica illumina. In questo senso, la storia selettiva è una *condizione di possibilità* del contenuto nei sistemi

biologici: spiega perché il sistema ha la struttura adatta a produrre rappresentazioni con contenuto.

Ma la determinazione del contenuto di ogni singola rappresentazione — il fatto che *questa* rappresentazione, prodotta *adesso* in risposta a *questi* stimoli, abbia *quel* contenuto — richiede una teoria supplementare che la teleosemantica non fornisce. È questa lacuna che la teoria del contenuto basata sulla codifica predittiva si propone di colmare, fondando la determinazione sincronica del contenuto sulla struttura del modello generativo e sul processo di minimizzazione dell'errore di predizione. Le due teorie non sono concorrenti ma complementari: la teleosemantica spiega la struttura del modello, la teoria predittiva spiega il contenuto delle rappresentazioni generate dal modello. Prima di sviluppare questa proposta, tuttavia, è necessario esaminare la seconda grande tradizione naturalistica del contenuto — la semantica informazionale di Dretske — per estrarne le risorse e identificarne i limiti specifici.

III — La semantica informazionale di Dretske: segnale, conoscenza e il problema della disgiuntività

III.1 — L'intuizione fondamentale: il contenuto come informazione trasportata

Fred Dretske ha elaborato la propria teoria del contenuto semantico a partire da una intuizione che è al tempo stesso semplice nella sua formulazione e ricca di conseguenze nel suo sviluppo: il contenuto di uno stato mentale è l'informazione che quello stato porta riguardo al mondo. Il riferimento alla teoria matematica dell'informazione di Shannon — che Dretske esplora con rigore in *Knowledge and the Flow of Information* (1981) — non è una metafora ma una proposta tecnica: l'informazione è una nozione oggettiva, misurabile, che descrive le relazioni di dipendenza probabilistica tra variabili, e la mente è un sistema che acquisisce, elabora e immagazzina informazione nel senso shannoniano del termine. Questa impostazione ha il pregio di ancorare il contenuto semantico a una nozione che non è già semantica — l'informazione è definita in termini puramente probabilistici, senza presupporre nozioni intenzionali — e di fornire criteri quantitativi per la determinazione del contenuto.

È necessario, prima di esaminare l'applicazione dretskiana al contenuto mentale, presentare con sufficiente precisione la struttura matematica della teoria shannoniana dell'informazione, poiché la comprensione precisa di quella struttura è indispensabile per valutare sia le risorse sia i limiti della semantica informazionale. Claude Shannon aveva sviluppato la propria teoria nel contesto dei

problemi ingegneristici della trasmissione del segnale attraverso canali rumorosi: il problema fondamentale era quantificare quanta informazione può essere trasmessa attraverso un canale di comunicazione con un certo rapporto segnale/rumore. L'informazione, in questo contesto, non ha nulla a che fare con il significato o il contenuto: è una misura puramente statistica della riduzione dell'incertezza prodotta dall'osservazione di un segnale.

Il punto di partenza tecnico è la definizione shannoniana di informazione. Dato un canale di comunicazione con sorgente S e ricevitore R , la quantità di informazione che un segnale r (un evento al ricevitore) porta riguardo alla sorgente s (un evento alla sorgente) è definita come la riduzione nell'incertezza circa s che si ottiene conoscendo r . Formalmente, se $P(s)$ è la probabilità a priori dell'evento s alla sorgente e $P(s|r)$ è la probabilità a posteriori di s dato r al ricevitore, la quantità di informazione trasportata da r riguardo a s è:

$$I(s; r) = \log_2 P(s|r) / P(s) = \log_2 P(s|r) - \log_2 P(s)$$

Quando r è tale che $P(s|r) = 1$ — quando la ricezione di r rende certa l'occorrenza di s alla sorgente — $I(s; r) = -\log_2 P(s)$, che è la quantità di informazione shannoniana associata all'evento s . Quando r è statisticamente indipendente da s , $P(s|r) = P(s)$ e $I(s; r) = 0$: r non porta alcuna informazione riguardo a s . La quantità di informazione è dunque una misura della dipendenza probabilistica tra r e s : quanta dell'incertezza circa s è ridotta dalla conoscenza di r .

L'entropia $H(S)$ di una sorgente con un insieme discreto di possibili eventi $\{s_1, s_2, \dots, s\}$ con probabilità $\{p_1, p_2, \dots, p\}$ è definita come il valore atteso dell'informazione di ciascun evento:

$$H(S) = -\sum_i p_i \log_2 p_i$$

L'entropia misura l'incertezza media della sorgente: una sorgente con distribuzione uniforme ha entropia massima ($\log_2 n$ per n eventi equiprobabili); una sorgente che produce sempre lo stesso evento ha entropia zero. L'informazione mutua $I(S; R)$ tra sorgente e ricevitore misura quanto, in media, la conoscenza del segnale al ricevitore riduce l'incertezza sulla sorgente:

$$I(S; R) = H(S) - H(S|R) = H(R) - H(R|S) = H(S) + H(R) - H(S, R)$$

dove $H(S|R)$ è l'entropia condizionale della sorgente dato il ricevitore, e $H(S, R)$ è l'entropia congiunta della sorgente e del ricevitore. L'informazione mutua è simmetrica e non-negativa, uguale a zero se e solo se S e R sono statisticamente indipendenti.

Dretske utilizza questa struttura matematica per definire il contenuto di un segnale, e poi, per estensione, il contenuto di uno stato cognitivo. Uno stato cognitivo C porta l'informazione che la proprietà F è istanziata nell'oggetto x se e solo se $P(F_x | C) = 1$ —

se la probabilità che x sia F è, dato C , uguale a uno. Il contenuto di C è dunque l'insieme delle proposizioni che C rende certe, nell'interpretazione più forte, o che C rende molto più probabili della loro probabilità a priori, in un'interpretazione più permissiva. L'informazione è la struttura invariante che si conserva attraverso le trasformazioni di segnale: se un segnale r_1 porta l'informazione che s , e r_1 causa r_2 in modo tale che la dipendenza tra r_2 ed s sia preservata, allora anche r_2 porta l'informazione che s .

III.2 — Dal segnale alla cognizione: l'acquisizione di conoscenza come flusso di informazione

Il passo cruciale nell'elaborazione di Dretske è il passaggio dall'informazione shannoniana — una nozione statistica, relazionale, senza implicazioni epistemiche o semantiche dirette — alla conoscenza e alla credenza. In *Knowledge and the Flow of Information*, Dretske propone che la conoscenza percettiva sia l'acquisizione di informazione da parte di un sistema cognitivo in un modo che influenza il sistema stesso. Un sistema K sa che s se e solo se K è in uno stato r che porta l'informazione che s , e questo stato è prodotto causalmente da s nel modo appropriato. La conoscenza è informazione che causa in modo appropriato stati nel conoscente.

La condizione che l'informazione sia *acquisita nel modo appropriato* è cruciale e ha una struttura che Dretske articola attraverso la nozione di *canale di informazione affidabile*. Un canale è affidabile se la dipendenza probabilistica tra la sorgente e il ricevitore è stabile attraverso una gamma sufficientemente ampia di condizioni: il canale uditivo che mi permette di percepire il suono di un campanello è affidabile se la probabilità che il campanello suoni, dato che ho la percezione uditiva corrispondente, è sufficientemente alta attraverso le condizioni normali di ascolto. L'affidabilità del canale è la condizione che trasforma la mera trasmissione di informazione in acquisizione di conoscenza. In simboli, se C_K è il canale che connette la sorgente s al sistema cognitivo K , e r è lo stato di K causato da s attraverso C_K , K sa che s se e solo se:

$$P(s \mid r, C_K \text{ affidabile}) = 1$$

La condizione $P(s \mid r) = 1$ — la certezza nomologica che r porta l'informazione che s — è la condizione di conoscenza piena; Dretske ammette versioni indebolite per la credenza giustificata, in cui la soglia probabilistica è inferiore a 1.

La nozione di canale affidabile introduce una componente normativa che non è presente nella pura teoria shannoniana dell'informazione. Shannon descriveva le proprietà statistiche di canali dati: la capacità di un canale è una proprietà matematica della distribuzione congiunta tra sorgente e ricevitore. Dretske non descrive canali dati ma prescrive condizioni di affidabilità: un canale è affidabile se mantiene le dipendenze probabilistiche

appropriate attraverso le condizioni normali di uso. Questa prescrizione introduce, surrettiziamente, una nozione normativa di «condizioni normali» che richiede una spiegazione: normali rispetto a quale standard? La risposta di Dretske oscilla tra un riferimento alle condizioni statisticamente prevalenti nell'ambiente del sistema — una risposta naturalistica che rischia la circolarità — e un riferimento alle condizioni in cui il sistema funziona come previsto dalla sua progettazione — una risposta che introduce considerazioni teleologiche.

III.3 — La transizione al contenuto semantico: la fase di digitalizzazione

Una delle contribuzioni più originali di Dretske alla teoria del contenuto è l'analisi della transizione tra l'elaborazione dell'informazione analogica — che caratterizza i livelli periferici del sistema sensoriale — e la rappresentazione digitale — che caratterizza i livelli superiori della cognizione. Dretske distingue tra *segnali analogici* e *segnali digitali* in senso tecnico. Un segnale r è una rappresentazione analogica di uno stato s se r porta informazione non solo su s ma anche su ogni proprietà di cui s è una determinazione: la sensazione visiva di un rosso scarlatta porta informazione non solo che la superficie è scarlatta ma che è rossa, colorata, avente riflettanza cromatica, ecc. Un segnale r è invece una rappresentazione digitale di s se porta esclusivamente l'informazione che s , ignorando le ulteriori specificazioni di s .

La digitalizzazione è il processo attraverso cui i sistemi cognitivi estrarrebbero informazione categoriale — digitale — da flussi di informazione continua — analogica. Il riconoscimento percettivo di un oggetto come appartenente a una categoria (questo è un gatto) è una forma di digitalizzazione: si estrae l'informazione categoriale (gatto) dall'informazione continua del flusso sensoriale, perdendo informazione sui dettagli specifici ma guadagnando in portabilità e in idoneità all'inferenza concettuale. Le credenze, nella teoria di Dretske, sono essenzialmente digitalizzazioni: sono stati che portano informazione categoriale estratta da flussi informativi più ricchi.

Formalmente, se il flusso analogico porta informazione sulla variabile continua X con distribuzione $P(X)$, la digitalizzazione produce una variabile discreta $D(X)$ che partizioni il dominio di X in categorie: $D(X) = d_i$ se $X \in R_i$, dove $\{R_i\}$ è una partizione del dominio. L'informazione portata dalla credenza digitale che $D(X) = d_i$ è:

$$I(D = d_i; r) = \log_2 P(D = d_i | r) / P(D = d_i)$$

La perdita di informazione nella digitalizzazione — l'informazione sui valori specifici di X all'interno di R_i — è compensata dalla maggiore trattabilità computazionale e dalla funzionalità inferenziale delle categorie digitali. La distinzione dretskiana tra elaborazione analogica e digitalizzazione categoriale anticipa in

modo sorprendentemente preciso la struttura gerarchica dei sistemi predittivi: i livelli inferiori della gerarchia predittiva elaborano segnali che assomigliano all'informazione analogica di Dretske, mentre i livelli superiori estraggono rappresentazioni categoriali che assomigliano alle sue digitalizzazioni.

C'è tuttavia una differenza cruciale tra la digitalizzazione dretskiana e la struttura gerarchica della codifica predittiva. Per Dretske, la digitalizzazione è un processo di riduzione: si perde informazione nell'estrarre la categoria dalla distribuzione continua. Per la codifica predittiva, la struttura gerarchica non è meramente riduttiva: i livelli superiori non si limitano a classificare gli input dei livelli inferiori, ma generano predizioni che scendono verso il basso attraverso la gerarchia e vincola attivamente l'elaborazione a tutti i livelli. La differenza è tra un flusso unidirezionale (analogico $\rightarrow$ digitale) e un flusso bidirezionale (predizioni discendenti $\leftrightarrow$ errori di predizione ascendenti). Questa bidirezionalità è, come si mostrerà nel Capitolo IV, la chiave per la soluzione del problema della disgiuntività.

III.4 — Il problema della disgiuntività: l'ostacolo principale alla semantica informazionale

Nonostante la sua eleganza tecnica, la semantica informazionale di Dretske affronta una difficoltà strutturale che ha occupato la letteratura per decenni e che non sembra avere una soluzione soddisfacente all'interno dei suoi presupposti. Questa difficoltà è nota come problema della disgiuntività, e la sua formulazione più precisa si deve a Fodor, anche se osservazioni analoghe erano state avanzate indipendentemente da diversi autori.

Il problema nasce dalla seguente osservazione. Il contenuto di uno stato cognitivo, secondo Dretske, è determinato dall'informazione che esso porta: lo stato C ha come contenuto «la proprietà F » se e solo se $P(F \mid C) = 1$ (o $P(F \mid C)$ è sufficientemente alta). Ma nella pratica dei sistemi cognitivi, uno stesso stato C può essere causato da più tipi di eventi: la mia percezione visiva di una forma felina può essere prodotta sia dalla presenza di un gatto sia da una sua rappresentazione realistica sia, in condizioni di scarsa illuminazione, da un oggetto di profilo simile. In ognuno di questi casi, $P(F_1 \vee F_2 \vee F_3 \mid C) = 1$, dove F_1 è la presenza di un gatto, F_2 è la presenza di una rappresentazione di un gatto, F_3 è la presenza di un oggetto con il profilo di un gatto. Ma allora il contenuto di C sarebbe la disgiuntiva $F_1 \vee F_2 \vee F_3$ — non «gatto», ma «gatto o rappresentazione di gatto o oggetto simile a un gatto» — e la misrappresentazione (prendere un gatto tassidermizzato per un gatto vivo) sarebbe impossibile: C avrebbe un contenuto disgiuntivo rispetto al quale è sempre corretta.

La struttura formale del problema può essere esposta in modo preciso. Sia l'insieme delle possibili cause di C il dominio $\{F_1, F_2, \dots, F\}$. La distribuzione di probabilità congiunta $P(C, F_1, F_2, \dots,$

F) è tale che $P(C \mid F_i) > 0$ per ogni i . L'informazione che C porta rispetto all'intera sorgente è:

$$I(C; F_1, F_2, \dots, F) = H(F_1, F_2, \dots, F) - H(F_1, F_2, \dots, F \mid C)$$

dove H indica l'entropia di Shannon. Questa quantità è ben definita e misura quanta incertezza sulla sorgente composita è ridotta dalla conoscenza di C . Ma per determinare il *contenuto specifico* di C — il fatto che C sia di F_1 piuttosto che di $F_1 \vee F_2$ — bisogna asimmetrizzare le dipendenze, individuare quale tra le F_i è la causa privilegiata. E questa asimmetrizzazione non è disponibile in termini puramente informativi: richiede risorse normative aggiuntive.

Dretske cerca di rispondere distinguendo tra le condizioni che determinano il contenuto di un segnale e le condizioni che interferiscono con la trasmissione del segnale. In un canale privo di rumore, C porta l'informazione che F_1 perché F_1 causa C attraverso il canale normale; F_2 e F_3 sono *sorgenti di rumore* — perturbazioni che possono produrre C anche in assenza di F_1 . La distinzione tra il segnale vero e il rumore presuppone però la nozione di *canale normale* o *condizioni normali di trasmissione*, che è a sua volta una nozione normativa. Come si determina quale sia il canale normale? In risposta, Dretske fa appello, nelle elaborazioni successive, a considerazioni teleologiche simili a quelle di Millikan: le condizioni normali sono quelle in cui il sistema informativo svolge la funzione per cui esiste o per cui è stato costruito. Ma questo reintroduce precisamente il tipo di fondamento teleologico che la semantica informativa intendeva evitare o supplementare.

La critica del disgiuntivista al programma informativo di Dretske mette in luce una tensione strutturale: la teoria dell'informazione di Shannon è una teoria dei canali di comunicazione, e la capacità di un canale è una proprietà globale che dipende da tutte le possibili sorgenti di segnale, non solo da quella «normale». Per determinare quale tra le possibili sorgenti è quella che «deve» causare il segnale — la sorgente che determina il contenuto piuttosto che il rumore — bisogna ricorrere a risorse normative che trascendono la struttura probabilistica del canale. La teoria dell'informazione, da sola, non fornisce queste risorse.

III.5 — La svolta in *Naturalizing the Mind*: sistemi di indicazione e funzioni di indicazione

Consapevole delle difficoltà del programma del 1981, Dretske ha significativamente riconfigurato la propria teoria in *Naturalizing the Mind* (1995), introducendo la nozione di *funzione di indicazione* — una nozione ibrida che combina l'informazione con considerazioni teleologiche. Un sistema S ha la funzione di indicare la proprietà F se S è stato reclutato — attraverso un processo di apprendimento o di selezione — per il compito di indicare F . La differenza rispetto alla teleosemantica di Millikan è che il reclutamento può avvenire

nell'arco della vita individuale attraverso il condizionamento e l'apprendimento, non solo attraverso la selezione filogenetica.

La struttura formale della funzione di indicazione è: S ha la funzione di indicare F se e solo se S è stato reclutato nella struttura del sistema cognitivo perché in passato l'indicazione di F da parte di S ha contribuito in modo causalmente rilevante al comportamento adattativo del sistema. Il contenuto di S è allora F — non la disgiuntiva delle sue cause possibili — perché è F la condizione che S deve indicare per assolvere la funzione per cui è stato reclutato.

Il processo di reclutamento che Dretske descrive nelle sue elaborazioni più dettagliate è modellato sul condizionamento operante: uno stato S viene «reclutato» come indicatore di F quando il comportamento prodotto in risposta a S (interpretato come indicatore di F) viene rinforzato selettivamente in presenza di F e non rinforzato in presenza delle cause spurie $F_2, F_3, \dots$. In questo modo, il sistema impara a trattare S come indicatore di F piuttosto che come indicatore della disgiuntiva, e questa storia di apprendimento differenziale è ciò che asimmetrizza le cause e determina il contenuto.

Questa revisione ha il merito di rispondere al problema della disgiuntività attraverso la storia di reclutamento: anche se S è causato sia da F_1 che da F_2 , il suo contenuto è F_1 se e solo se S è stato reclutato perché indicava F_1 , non perché indicava F_2 . Ma al prezzo di reintrodurre una dimensione temporale e teleologica che avvicina la teoria di Dretske a quella di Millikan, riducendo la novità della proposta informazionale. Il contributo specifico della semantica informazionale — la formalizzazione del contenuto in termini di dipendenze probabilistiche, senza fare appello alla storia selettiva — viene in parte sacrificato sull'altare della risposta al problema della disgiuntività.

C'è un'altra difficoltà che la revisione del 1995 introduce senza risolvere: il problema del cerchio del reclutamento. Perché un sistema apprende a trattare S come indicatore di F piuttosto che di $F_1 \vee F_2$? La risposta del condizionamento è che il comportamento in risposta a S viene rinforzato quando F è presente e non quando non lo è. Ma questa risposta presuppone che il sistema possa distinguere le situazioni in cui F è presente da quelle in cui non lo è — il che presuppone già la capacità di rappresentare F, non solo di rispondere a stimoli che covariano con F. Il reclutamento sembra presupporre già la capacità di rappresentare il contenuto che il reclutamento dovrebbe fondare, generando una circolarità che è difficile da evitare.

III.6 — I contributi permanenti e i limiti della semantica informazionale

Il bilancio dell'approccio informazionale di Dretske, ai fini della presente indagine, può essere articolato in tre punti. Il contributo

più durevole è la formalizzazione esplicita della relazione tra stati cognitivi e stati del mondo in termini probabilistici: l'idea che il contenuto sia determinato dalla struttura delle dipendenze statistiche tra il sistema e l'ambiente è — come si argomenterà — un'intuizione fondamentale che il framework della codifica predittiva riprende e trasforma in modo sostanziale. La differenza cruciale è che nella codifica predittiva le dipendenze statistiche non fluiscono unicamente dalla sorgente al ricevitore lungo canali passivi: il sistema cognitivo costruisce attivamente un modello generativo che *genera* predizioni su quelle dipendenze, e l'informazione che determina il contenuto è l'informazione modellata, non quella semplicemente trasmessa.

Il secondo contributo rilevante è la distinzione tra elaborazione analogica e digitalizzazione categoriale, che anticipa in modo sorprendentemente preciso la struttura gerarchica dei sistemi predittivi. I livelli inferiori della gerarchia predittiva elaborano segnali che assomigliano all'informazione analogica di Dretske, mentre i livelli superiori estraggono rappresentazioni categoriali che assomigliano alle sue digitalizzazioni. La differenza è che nella codifica predittiva le due modalità di elaborazione non sono separate in due fasi sequenziali ma interagiscono bidirezionalmente attraverso i segnali di errore di predizione.

Il terzo contributo, spesso sottovalutato, è la connessione tra il contenuto semantico e la struttura della conoscenza percettiva. Dretske ha sostenuto in modo convincente che la conoscenza percettiva è qualcosa di più della semplice registrazione causale di fatti: è l'acquisizione di informazione attraverso canali affidabili, e l'affidabilità del canale è una proprietà del sistema cognitivo nel suo complesso, non solo del singolo stato percettivo. Questa prospettiva sistemica sulla conoscenza è coerente con l'approccio della codifica predittiva, in cui il contenuto di ogni rappresentazione dipende dalla struttura dell'intero modello generativo e non solo dalla relazione tra il singolo stato e il suo stimolo.

Il limite principale, che si aggiunge a quello della disgiuntività già esaminato, è la passività strutturale della semantica informazionale: il sistema cognitivo è concepito essenzialmente come un ricevitore di informazione, non come un generatore attivo di predizioni. Questa passività va contro l'evidenza neurobiologica e computazionale che indica la cognizione come un processo fondamentalmente predittivo, top-down, in cui il sistema non si limita a ricevere input ma genera attivamente aspettative che strutturano l'elaborazione dell'input. È questa attività generativa che la teoria del contenuto basata sulla codifica predittiva intende porre al centro, e che richiede un apparato formale diverso da quello che la semantica informazionale mette a disposizione.

Un'ultima considerazione riguarda il rapporto tra la semantica informazionale e il problema della coscienza fenomenica. Dretske, nelle sue elaborazioni su questo tema (in particolare in *Naturalizing*

the Mind), ha sostenuto che la coscienza di un evento E è l'avere uno stato che porta l'informazione che E è presente, dove il portare è inteso come una proprietà intrinseca dello stato, non come una relazione con un interprete. Questa proposta tenta di ridurre la coscienza alla proprietà informazionale, ma è difficile evitare la seguente obiezione: i sistemi di elaborazione dell'informazione non fenomenicamente coscienti — come i circuiti elettronici — portano anch'essi informazione in senso shannoniano, e dunque la proprietà informazionale da sola non può essere la condizione sufficiente della coscienza. La risposta di Dretske — che la proprietà rilevante non è il portare informazione in senso shannoniano ma il *misrepresentare* in modo appropriato, e che questo richiede determinate strutture funzionali che i circuiti semplici non hanno — lascia aperta la domanda su quali strutture funzionali siano necessarie e perché quelle strutture producano coscienza. Il problema difficile della coscienza rimane, anche all'interno della semantica informazionale.

IV — I fondamenti matematici della codifica predittiva gerarchica: dal principio di energia libera alle reti predittive

IV.1 — Il problema della percezione come inferenza bayesiana

La codifica predittiva gerarchica non è una teoria filosofica del contenuto semantico ma un framework computazionale e neurobiologico che descrive come il cervello elabora l'informazione sensoriale. La sua rilevanza per la filosofia del contenuto deriva dal fatto che esso offre, per la prima volta in forma matematicamente precisa, una caratterizzazione del processo attraverso cui un sistema cognitivo genera rappresentazioni con contenuto determinato — una caratterizzazione che ha implicazioni dirette sui problemi che le teorie precedenti affrontavano in termini qualitativi. Prima di sviluppare queste implicazioni è indispensabile esporre con rigore la struttura matematica del framework, poiché le conclusioni filosofiche dipendono in modo essenziale dai dettagli formali.

Il punto di partenza è la rilettura della percezione come problema di inferenza statistica. Il problema della percezione, in termini generali, è il problema inverso: dato un flusso di segnali sensoriali $\mathbf{o}$ (osservazioni), inferire le cause $\mathbf{s}$ di quei segnali — le proprietà e gli stati dell'ambiente esterno che hanno generato i segnali osservati. Questo problema è mal posto nel senso matematico di Hadamard: lo stesso insieme di osservazioni $\mathbf{o}$ è compatibile in linea di principio con una molteplicità di stati causali $\mathbf{s}$, e non esiste una procedura algebricamente semplice per invertire il processo causale che va da $\mathbf{s}$ a $\mathbf{o}$. La percezione come inversione del

problema causale richiede un vincolo aggiuntivo che riduca lo spazio delle soluzioni compatibili con le osservazioni.

Il carattere mal posto del problema della percezione non è un difetto delle teorie matematiche ma una caratteristica strutturale della situazione epistemica di qualsiasi sistema cognitivo incarnato nel mondo. Lo stesso oggetto fisico può apparire in migliaia di modi diversi a seconda delle condizioni di illuminazione, della posizione del punto di vista, delle occlusioni parziali, della distanza. Il sistema visivo deve ricostruire l'oggetto tridimensionale da proiezioni bidimensionali ambigue. Helmholtz aveva già compreso questa struttura nel XIX secolo con la nozione di «inferenza inconscia» (*unbewusste Schluss*): la percezione non è una registrazione passiva ma un processo attivo di inferenza, in cui le sensazioni presenti vengono interpretate alla luce di conoscenze e aspettative precedentemente acquisite. La rivoluzione bayesiana nella percezione non fa che formalizzare matematicamente questa intuizione helmholtziana.

Il framework bayesiano fornisce il vincolo necessario attraverso la distribuzione a priori sugli stati causali: $P(\mathbf{s})$, la distribuzione di probabilità che il sistema assegna agli stati causali prima di ricevere qualsiasi osservazione. La distribuzione a posteriori — la distribuzione sugli stati causali dopo aver osservato $\mathbf{o}$ — è data dalla regola di Bayes:

$$P(\mathbf{s} \mid \mathbf{o}) = P(\mathbf{o} \mid \mathbf{s}) \cdot P(\mathbf{s}) / P(\mathbf{o})$$

dove $P(\mathbf{o} \mid \mathbf{s})$ è la funzione di verosimiglianza — la probabilità di osservare $\mathbf{o}$ dato che lo stato causale è $\mathbf{s}$ — e $P(\mathbf{o}) = \int P(\mathbf{o} \mid \mathbf{s}) P(\mathbf{s}) d\mathbf{s}$ è la probabilità marginale delle osservazioni, che funge da costante di normalizzazione. La percezione ottimale consiste nel calcolare la distribuzione a posteriori $P(\mathbf{s} \mid \mathbf{o})$ e derivarne una stima degli stati causali — tipicamente la moda della distribuzione (stima MAP, Maximum A Posteriori) o il valore atteso $E[\mathbf{s} \mid \mathbf{o}]$.

La regola di Bayes cattura l'essenza del processo percettivo ottimale: le credenze a posteriori sugli stati del mondo bilanciano le informazioni a priori (cosa ci si aspettava prima di osservare l'input) con le informazioni fornite dall'input (quanto l'input è compatibile con diversi stati del mondo). Il denominatore $P(\mathbf{o})$ garantisce la normalizzazione ma è tipicamente insolubile analiticamente per sistemi reali: richiede l'integrazione su tutti i possibili stati causali, che in sistemi ad alta dimensionalità è computazionalmente intrattabile. Questa intrattabilità è la radice matematica della necessità di metodi approssimati come l'inferenza variazionale.

La separazione tra prior e likelihood — le due componenti della regola di Bayes — ha una risonanza diretta con il dibattito tra internalismo ed externalismo nella filosofia del contenuto. Il prior codifica le credenze del sistema indipendentemente dall'input corrente: è la componente «interna» nel senso funzionalista. La

likelihood codifica le relazioni tra gli stati causali e gli input osservabili: è la componente «esterna» nel senso che riflette la struttura fisica del mondo e le proprietà causali degli stati causali rispetto ai segnali sensoriali. Il processo di inferenza bayesiana è un processo di integrazione della componente interna e di quella esterna, e il contenuto delle rappresentazioni risultanti è determinato da questa integrazione. La distinzione interno/esterno non scompare nel framework bayesiano, ma viene ricollocata: non è una distinzione tra tipi di sistemi (internalisti vs. externalisti) ma tra componenti dello stesso processo (prior vs. likelihood).

IV.2 — L'inferenza variazionale e l'approssimazione della distribuzione a posteriori

Il problema fondamentale dell'inferenza bayesiana applicata a sistemi cognitivi reali è la trattabilità computazionale. Per spazi di stati causali ad alta dimensionalità — che è la situazione normale per qualsiasi sistema cognitivo biologicamente plausibile — il calcolo diretto della distribuzione a posteriori richiede l'integrazione su uno spazio di enorme dimensionalità, che è computazionalmente intrattabile con le risorse disponibili a sistemi neurali o artificiali. Il framework della codifica predittiva risolve questo problema attraverso un'approssimazione variazionale: invece di calcolare la vera distribuzione a posteriori, il sistema mantiene una distribuzione approssimante $Q(\mathbf{s}; \boldsymbol{\lambda})$ — parametrizzata da $\boldsymbol{\lambda}$ — e ottimizza i parametri $\boldsymbol{\lambda}$ in modo da rendere Q il più vicina possibile alla vera posteriori $P(\mathbf{s} \mid \mathbf{o})$.

L'approccio variazionale ha radici nella fisica statistica, dove la nozione di energia libera di Helmholtz era già usata nel XIX secolo per approssimare le proprietà di sistemi fisici complessi. L'idea di base è che invece di calcolare esattamente la funzione di partizione — il denominatore della distribuzione di Boltzmann — si introduce una distribuzione di prova parametrizzata e si ottimizzano i parametri per minimizzare la divergenza dalla vera distribuzione. Questo approccio è computazionalmente trattabile nella misura in cui la famiglia di distribuzioni di prova è sufficientemente semplice, e produce approssimazioni utili nella misura in cui la famiglia è sufficientemente ricca da catturare le caratteristiche principali della vera distribuzione.

Nel contesto dell'inferenza bayesiana per la percezione, l'approssimazione variazionale procede come segue. Si sceglie una famiglia di distribuzioni parametrizzate $Q(\mathbf{s}; \boldsymbol{\lambda})$ — tipicamente, il campo medio, ovvero il prodotto di distribuzioni indipendenti su ciascuna variabile latente: $Q(\mathbf{s}; \boldsymbol{\lambda}) = \prod_i Q_i(s_i; \lambda_i)$. Si definisce la misura di distanza tra Q e la vera posteriori $P(\mathbf{s} \mid \mathbf{o})$ come la divergenza di Kullback-Leibler:

$$D_{\text{KL}}[Q(\mathbf{s}; \boldsymbol{\lambda}) \parallel P(\mathbf{s} \mid \mathbf{o})] = \int Q(\mathbf{s}; \boldsymbol{\lambda}) \log [Q(\mathbf{s}; \boldsymbol{\lambda}) / P(\mathbf{s} \mid \mathbf{o})] d\mathbf{s}$$

La divergenza KL è non-negativa e uguale a zero se e solo se $Q(\mathbf{s}; \boldsymbol{\lambda}) = P(\mathbf{s} \mid \mathbf{o})$ quasi ovunque. Minimizzare D_{KL} rispetto a $\boldsymbol{\lambda}$

equivale dunque a trovare la distribuzione approssimante che meglio approssima la vera posteriori.

Il problema computazionale è che la vera posteriori $P(\mathbf{s} \mid \mathbf{o})$ non è direttamente calcolabile — è quella la quantità che stiamo cercando. Si introduce allora l'*energia libera variazionale* $\mathbf{F}$, definita come:

$$\mathbf{F}[Q, \mathbf{o}] = D_{\text{KL}}[Q(\mathbf{s}; \boldsymbol{\lambda}) \parallel P(\mathbf{s} \mid \mathbf{o})] - \log P(\mathbf{o})$$

Poiché $\log P(\mathbf{o})$ non dipende da Q , minimizzare $\mathbf{F}$ rispetto a $\boldsymbol{\lambda}$ è equivalente a minimizzare $D_{\text{KL}}[Q \parallel P(\mathbf{s} \mid \mathbf{o})]$. Ma l'energia libera ha il vantaggio di essere calcolabile direttamente, senza conoscere $P(\mathbf{s} \mid \mathbf{o})$. La derivazione esplicita è la seguente. Sostituendo la definizione di $P(\mathbf{s} \mid \mathbf{o}) = P(\mathbf{o} \mid \mathbf{s}) \cdot P(\mathbf{s}) / P(\mathbf{o})$:

$$\begin{aligned} \mathbf{F}[Q, \mathbf{o}] &= \int Q(\mathbf{s}; \boldsymbol{\lambda}) \log [Q(\mathbf{s}; \boldsymbol{\lambda}) / (P(\mathbf{o} \mid \mathbf{s}) \cdot P(\mathbf{s}) / P(\mathbf{o}))] d\mathbf{s} - \log P(\mathbf{o}) \\ &= \int Q(\mathbf{s}; \boldsymbol{\lambda}) \log [Q(\mathbf{s}; \boldsymbol{\lambda}) \cdot P(\mathbf{o}) / (P(\mathbf{o} \mid \mathbf{s}) \cdot P(\mathbf{s}))] d\mathbf{s} - \log P(\mathbf{o}) \\ &= \int Q(\mathbf{s}; \boldsymbol{\lambda}) [\log Q(\mathbf{s}; \boldsymbol{\lambda}) - \log P(\mathbf{s}) - \log P(\mathbf{o} \mid \mathbf{s}) + \log P(\mathbf{o})] d\mathbf{s} - \log P(\mathbf{o}) \\ &= \int Q(\mathbf{s}; \boldsymbol{\lambda}) \log [Q(\mathbf{s}; \boldsymbol{\lambda}) / P(\mathbf{s})] d\mathbf{s} - \int Q(\mathbf{s}; \boldsymbol{\lambda}) \log P(\mathbf{o} \mid \mathbf{s}) d\mathbf{s} \\ &= D_{\text{KL}}[Q(\mathbf{s}; \boldsymbol{\lambda}) \parallel P(\mathbf{s})] - E_Q[\log P(\mathbf{o} \mid \mathbf{s})] \end{aligned}$$

L'energia libera si decompone in due termini: il primo, $D_{\text{KL}}[Q \parallel P(\mathbf{s})]$, è la divergenza KL tra la distribuzione approssimante e la distribuzione a priori — misura quanto Q si discosta dalle credenze a priori del sistema; il secondo, $-E_Q[\log P(\mathbf{o} \mid \mathbf{s})]$, è il valore atteso negativo del logaritmo della verosimiglianza — misura quanto bene la distribuzione approssimante rende conto delle osservazioni. La minimizzazione dell'energia libera bilancia questi due termini: il sistema cerca la distribuzione sugli stati causali che sia al tempo stesso vicina alla priori e coerente con le osservazioni. Questo bilanciamento è la forma matematica dell'equilibrio epistemico tra credenze preesistenti e nuove evidenze.

IV.3 — Le tre decomposizioni dell'energia libera e le loro interpretazioni filosofiche

L'energia libera variazionale ammette tre decomposizioni equivalenti che illuminano aspetti diversi del processo cognitivo e hanno implicazioni filosofiche distinte. È utile presentarle tutte e tre, poiché ciascuna mette in risalto una dimensione diversa della struttura della cognizione come minimizzazione dell'energia libera.

La prima decomposizione è quella già presentata:

$$\mathbf{F}[Q, \mathbf{o}] = D_{\text{KL}}[Q(\mathbf{s}; \boldsymbol{\lambda}) \parallel P(\mathbf{s})] - E_Q[\log P(\mathbf{o} \mid \mathbf{s})]$$

In questa forma, l'energia libera è la differenza tra la *complessità* della distribuzione approssimante — quanto essa diverge dalla priori, che misura il costo in termini di credenze precedenti di adottare Q come modello del mondo — e l'*accuratezza* del modello — quanto bene il modello rende conto delle osservazioni. La minimizzazione dell'energia libera bilancia complessità e accuratezza: il sistema cerca il modello più semplice compatibile con i dati, in accordo con il principio di parsimonia epistemica (il rasoio di Occam bayesiano). Questa decomposizione ha una risonanza diretta con la nozione di modelli generativi nell'apprendimento automatico: un buon modello è quello che comprime efficientemente i dati (bassa complessità) mantenendo un'alta accuratezza nella loro riproduzione.

La seconda decomposizione è:

$$\mathbf{F}[Q, \mathbf{o}] = -H[Q] - E_Q[\log P(\mathbf{o}, \mathbf{s})]$$

dove $H[Q] = -\int Q \log Q \, d\mathbf{s}$ è l'entropia della distribuzione approssimante. In questa forma, minimizzare $\mathbf{F}$ equivale a massimizzare la somma dell'entropia di Q e del valore atteso del logaritmo della probabilità congiunta $P(\mathbf{o}, \mathbf{s})$. Il termine entropico favorisce distribuzioni approssimanti più uniformi — distribuzioni che mantengono incertezza sugli stati causali — mentre il termine di log-verosimiglianza congiunta favorisce distribuzioni che assegnano alta probabilità agli stati causali che meglio spiegano le osservazioni congiuntamente alla priori. L'interpretazione filosofica di questa decomposizione riguarda il rapporto tra certezza e umiltà epistemica: il sistema non deve essere né troppo certo (bassa entropia, rigidità del modello) né troppo incerto (alta entropia, nessuna struttura cognitiva). L'ottimale è il punto di massima entropia compatibile con i dati, che nella letteratura sull'inferenza bayesiana è noto come principio di massima entropia di Jaynes.

La terza decomposizione è quella più direttamente rilevante per il framework della codifica predittiva:

$$\mathbf{F}[Q, \mathbf{o}] = -\log P(\mathbf{o}) + D_{\text{KL}}[Q(\mathbf{s}; \lambda) \parallel P(\mathbf{s} \mid \mathbf{o})]$$

Poiché $D_{\text{KL}} \geq 0$, ne segue che $\mathbf{F}[Q, \mathbf{o}] \geq -\log P(\mathbf{o})$. L'energia libera è un *upper bound* sulla sorpresa sensoriale $-\log P(\mathbf{o})$. Minimizzare l'energia libera equivale a minimizzare un limite superiore sulla sorpresa. Questa è l'affermazione centrale del principio di energia libera di Friston: i sistemi biologici autoorganizzanti si comportano come se minimizzassero l'energia libera variazionale, il che equivale ad approssimare l'inferenza bayesiana ottimale minimizzando allo stesso tempo la sorpresa per le loro osservazioni sensoriali.

La sorpresa $-\log P(\mathbf{o})$ ha un'interpretazione fisica ed evolutiva importante. Un sistema biologico che riceve osservazioni con alta sorpresa — osservazioni molto improbabili rispetto al suo modello — si trova in una situazione lontana dalla configurazione che il suo

modello prevede come normale. Se la sorpresa è sistematicamente alta, il sistema sta operando in un regime che il suo modello non riesce a predire, il che è incompatibile con la sopravvivenza a lungo termine. I sistemi biologici che sopravvivono sono quelli che costruiscono modelli accurati dell'ambiente, riducendo la sorpresa e operando in regimi che i loro modelli predicono. La minimizzazione della sorpresa — ovvero dell'energia libera — è dunque il correlato computazionale dell'imperativo evolutivo di sopravvivere e riprodursi.

IV.4 — La struttura gerarchica: modelli generativi a livelli multipli

Fino a questo punto, il framework è stato presentato per sistemi con un solo livello di rappresentazione. La codifica predittiva gerarchica generalizza questa struttura a sistemi con una gerarchia di livelli di rappresentazione, in cui ciascun livello genera predizioni per il livello inferiore e riceve segnali di errore di predizione da quel medesimo livello. Questa generalizzazione è cruciale sia sul piano neurobiologico — le aree corticali sono organizzate gerarchicamente, con connessioni discendenti che trasportano predizioni e connessioni ascendenti che trasportano errori di predizione — sia sul piano filosofico, poiché la struttura gerarchica è essenziale per la determinazione del contenuto semantico a diversi livelli di astrazione.

Sia una gerarchia di L livelli, con $\mathbf{s}^{(l)}$ che denota gli stati rappresentazionali al livello l , per $l = 1, \dots, L$. Le osservazioni sensoriali $\mathbf{o}$ si trovano al livello 0. Il modello generativo gerarchico specifica una distribuzione congiunta sulle osservazioni e su tutti i livelli di stati causali:

$$P(\mathbf{o}, \mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \dots, \mathbf{s}^{(L)}) = P(\mathbf{o} \mid \mathbf{s}^{(1)}) \cdot \prod_{l=1}^{L-1} P(\mathbf{s}^{(l)} \mid \mathbf{s}^{(l+1)}) \cdot P(\mathbf{s}^{(L)})$$

dove $P(\mathbf{o} \mid \mathbf{s}^{(1)})$ è la verosimiglianza delle osservazioni dati gli stati al primo livello, $P(\mathbf{s}^{(l)} \mid \mathbf{s}^{(l+1)})$ è la distribuzione condizionale degli stati al livello l dati gli stati al livello $l+1$, e $P(\mathbf{s}^{(L)})$ è la distribuzione a priori al livello più alto. Questa struttura fattorizzata è la *Markov property* gerarchica: dato lo stato al livello $l+1$, lo stato al livello l è condizionalmente indipendente dagli stati a tutti i livelli superiori.

Il modello generativo gerarchico può essere scritto in termini di funzioni di mapping tra livelli. Definiamo $\mathbf{g}^{(l)}(\mathbf{s}^{(l+1)})$ come la funzione che mappa gli stati al livello $l+1$ nelle previsioni degli stati al livello l , e $\mathbf{f}^{(l)}(\mathbf{s}^{(l)}, \mathbf{v}^{(l)})$ come la funzione di dinamica che descrive l'evoluzione degli stati al livello l in risposta ai drive esterni $\mathbf{v}^{(l)}$. La struttura generativa assume la forma di equazioni di stato annidate:

$$\mathbf{s}^{(l-1)} = \mathbf{g}^{(l)}(\mathbf{s}^{(l)}) + \boldsymbol{\epsilon}^{(l)}$$

dove $\boldsymbol{\epsilon}^{(l)}$ è il rumore a livello l , tipicamente modellato come gaussiano con precisione $\Pi^{(l)}$ (l'inverso della covarianza):

$$\boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(0, (\Pi^{(l)})^{-1})$$

In questa formulazione, il modello generativo è un sistema di equazioni differenziali stocastiche annidate che descrivono come gli stati ai livelli superiori *generano* gli stati ai livelli inferiori attraverso le funzioni $\mathbf{g}^{(l)}$, con rumore additivo gaussiano. Il termine «generativo» riflette il fatto che il modello può essere usato per *generare* campioni delle osservazioni sensoriali *prima* di riceverle, producendo predizioni top-down. È questa capacità generativa — la possibilità di usare il modello nella direzione opposta rispetto al flusso causale naturale — che distingue strutturalmente il framework predittivo dalla semantica informazionale passiva di Dretske.

La scelta di modellare il rumore come gaussiano è motivata sia da considerazioni matematiche — le gaussiane sono la famiglia di distribuzioni che massimizza l'entropia sotto il vincolo di media e varianza date, e hanno proprietà di chiusura convenienti sotto la convoluzione e la somma — sia da evidenze neuroscientifiche che suggeriscono distribuzioni approssimativamente gaussiane per il rumore neurale a diversi livelli della gerarchia corticale. Nella forma gaussiana, le densità di probabilità ai diversi livelli sono:

$$P(\mathbf{s}^{(l-1)} | \mathbf{s}^{(l)}) \propto \exp(-1/2 (\mathbf{s}^{(l-1)} - \mathbf{g}^{(l)}(\mathbf{s}^{(l)}))^T \Pi^{(l)} (\mathbf{s}^{(l-1)} - \mathbf{g}^{(l)}(\mathbf{s}^{(l)})))$$

e la log-verosimiglianza condizionale è la forma quadratica pesata per la precisione:

$$\log P(\mathbf{s}^{(l-1)} | \mathbf{s}^{(l)}) = -1/2 (\mathbf{s}^{(l-1)} - \mathbf{g}^{(l)}(\mathbf{s}^{(l)}))^T \Pi^{(l)} (\mathbf{s}^{(l-1)} - \mathbf{g}^{(l)}(\mathbf{s}^{(l)})) + \text{costante}$$

La trattabilità matematica delle gaussiane rende possibile derivare formule analitiche per l'energia libera e le equazioni di aggiornamento, che nel caso non-gaussiano richiederebbero metodi numerici o approssimazioni ulteriori.

IV.5 — La minimizzazione dell'energia libera nella gerarchia: gli errori di predizione pesati per la precisione

L'energia libera variazionale per il modello gerarchico, con distribuzione approssimante $Q = \prod Q^{(l)}(\mathbf{s}^{(l)}; \boldsymbol{\mu}^{(l)})$ — il campo medio, con parametri variazionali $\boldsymbol{\mu}^{(l)}$ che rappresentano le credenze a posteriori agli stati causali di ciascun livello — ha la forma:

$$\mathbf{F} = \sum \{ 1/2 (\epsilon_l^{(l)})^T \Pi^{(l)} \epsilon_l^{(l)} - 1/2 \log |\Pi^{(l)}| + \text{costanti} \}$$

dove $\epsilon_l^{(l)} = \boldsymbol{\mu}^{(l-1)} - \mathbf{g}^{(l)}(\boldsymbol{\mu}^{(l)})$ è l'errore di predizione al livello l — la discrepanza tra lo stato stimato al livello $l-1$ e la predizione di quello stato generata dallo stato stimato al livello l attraverso la funzione $\mathbf{g}^{(l)}$.

Le condizioni di punto stazionario per la minimizzazione di $\mathbf{F}$ rispetto ai parametri variazionali $\boldsymbol{\mu}^{(l)}$ si ottengono imponendo $\partial \mathbf{F} / \partial \boldsymbol{\mu}^{(l)} = 0$. Il calcolo esplicito richiede di differenziare l'energia

libera rispetto a $\mu^{(l)}$, tenendo conto che $\mu^{(l)}$ appare sia come argomento di $\mathbf{g}^{(l)}$ (nei termini del livello l) sia come argomento dell'errore $\epsilon_e^{(l+1)} = \mu^{(l)} - \mathbf{g}^{(l+1)}(\mu^{(l+1)})$ (nei termini del livello $l+1$):

$$\partial \mathbf{F} / \partial \mu^{(l)} = -(\partial \mathbf{g}^{(l)} / \partial \mu^{(l)})^T \Pi^{(l)} \epsilon_e^{(l)} + \Pi^{(l+1)} \epsilon_e^{(l+1)} = 0$$

dove il primo termine rappresenta la derivata dell'errore al livello l rispetto alle credenze al livello l (propagata attraverso la jacobiana $\partial \mathbf{g}^{(l)} / \partial \mu^{(l)}$), e il secondo termine è l'errore pesato al livello $l+1$ — l'errore tra le credenze al livello l e la predizione generata dal livello $l+1$. In forma compatta, le condizioni di stazionarietà sono:

$$\Pi^{(l+1)} \epsilon_e^{(l+1)} = (\partial \mathbf{g}^{(l)} / \partial \mu^{(l)})^T \Pi^{(l)} \epsilon_e^{(l)}$$

Questa equazione ha una lettura straordinariamente elegante: le credenze a ciascun livello sono aggiornate fino al punto in cui l'errore di predizione pesato proveniente dal livello inferiore — il lato destro, che descrive quanto le credenze attuali al livello l non riescono a predire le credenze al livello $l-1$ — è bilanciato dall'errore di predizione pesato che queste credenze generano verso il livello superiore — il lato sinistro, che descrive quanto le credenze al livello l non si accordano con le predizioni del livello $l+1$. All'equilibrio, ogni livello è simultaneamente consistente con le predizioni che riceve dall'alto e con le osservazioni che riceve dal basso: il sistema si trova in una configurazione di minima tensione epistemica.

Per sistemi dinamici in cui gli stati evolvono nel tempo, la minimizzazione dell'energia libera si realizza come un sistema di equazioni differenziali di gradiente-discesa sui parametri variazionali:

$$\dot{\mu}^{(l)} = -\partial \mathbf{F} / \partial \mu^{(l)} = (\partial \mathbf{g}^{(l)} / \partial \mu^{(l)})^T \Pi^{(l)} \epsilon_e^{(l)} - \Pi^{(l+1)} \epsilon_e^{(l+1)}$$

Il sistema evolve lungo il gradiente negativo dell'energia libera fino a raggiungere un punto stazionario — la stima a posteriori ottimale degli stati causali. Questo sistema di equazioni differenziali descrive l'aggiornamento delle rappresentazioni interne del sistema cognitivo in risposta agli errori di predizione: è la dinamica dell'inferenza percettiva in tempo reale.

L'interpretazione neurobiologica di queste equazioni è che ciascun neurone corticale implementa una piccola parte del calcolo: i neuroni nelle proiezioni ascendenti (feedforward) trasmettono gli errori di predizione pesati, i neuroni nelle proiezioni discendenti (feedback) trasmettono le predizioni generate dai livelli superiori, e i neuroni che sommano questi segnali calcolano le nuove stime dei parametri variazionali. Il risultato è un sistema massivamente parallelo di calcolo distribuito che realizza l'inferenza bayesiana approssimata senza nessun centro di controllo, emergendo dall'interazione locale di miliardi di neuroni.

IV.6 — L'approssimazione di Laplace e le soluzioni analitiche

Per il caso in cui la distribuzione a posteriori è approssimativamente gaussiana — un'assunzione ragionevole per molti problemi percettivi — è possibile applicare l'approssimazione di Laplace, che fornisce soluzioni analitiche esplicite per le equazioni di aggiornamento. L'approssimazione di Laplace consiste nell'espandere il logaritmo della distribuzione a posteriori attorno alla sua moda $\mu^{(l)}$ a secondo ordine:

$$\log P(\mathbf{s}^{(l)} \mid \mathbf{o}, \mathbf{s}^{(l+1)}) \approx \log P(\mu^{(l)} \mid \mathbf{o}, \mathbf{s}^{(l+1)}) - \frac{1}{2} (\mathbf{s}^{(l)} - \mu^{(l)})^T \Lambda^{(l)} (\mathbf{s}^{(l)} - \mu^{(l)})$$

dove $\Lambda^{(l)} = -\partial^2 \log P(\mathbf{s}^{(l)} \mid \mathbf{o}, \mathbf{s}^{(l+1)}) / \partial \mathbf{s}^{(l)2} \big|_{\{\mu^{(l)}\}}$ è la matrice hessiana negativa del logaritmo della a posteriori valutata alla moda. Nell'approssimazione di Laplace, la distribuzione approssimante $Q^{(l)}$ è una gaussiana con media $\mu^{(l)}$ e matrice di covarianza $(\Lambda^{(l)})^{-1}$:

$$Q^{(l)}(\mathbf{s}^{(l)}) = \mathcal{N}(\mu^{(l)}, (\Lambda^{(l)})^{-1})$$

La matrice hessiana $\Lambda^{(l)}$ combina la curvatura della verosimiglianza e quella della priori:

$$\begin{aligned} \Lambda^{(l)} &= (\partial^2 / \partial \mathbf{s}^{(l)2}) [-\log P(\mathbf{s}^{(l)} \mid \mathbf{s}^{(l+1)}) - \log P(\mathbf{s}^{(l)} \mid \mathbf{s}^{(l+1)})] \\ &= (\partial \mathbf{g}^{(l)} / \partial \mathbf{s}^{(l)})^T \Pi^{(l)} (\partial \mathbf{g}^{(l)} / \partial \mathbf{s}^{(l)}) + \Pi^{(l+1)} \end{aligned}$$

Questa espressione mostra come la precisione a posteriori al livello l è la somma delle contribuzioni della verosimiglianza (prima riga, mediata attraverso la jacobiana della funzione generativa) e della priori (seconda riga, la precisione al livello $l+1$). La precisione a posteriori è sempre maggiore o uguale alla precisione della priori, riflettendo il fatto che l'osservazione riduce l'incertezza.

L'approssimazione di Laplace è particolarmente utile per il calcolo dell'energia libera in forma chiusa. Nell'approssimazione gaussiana:

$$\mathbf{F}_{\text{Laplace}} = \frac{1}{2} \sum [(\epsilon_e^{(l)})^T \Pi^{(l)} \epsilon_e^{(l)} - \log |\Pi^{(l)}| + n_l \log(2\pi)]$$

dove n_l è la dimensionalità dello stato al livello l , e l'errore di predizione $\epsilon_e^{(l)} = \mu^{(l+1)} - \mathbf{g}^{(l)}(\mu^{(l)})$ è valutato alla stima MAP. In questa forma, l'energia libera è la somma degli errori di predizione quadratici pesati per la precisione a tutti i livelli, corretta per i termini logaritmici della normalizzazione gaussiana. Questa è la quantità che il sistema cognitivo minimizza attraverso le equazioni di aggiornamento: ogni aggiornamento delle credenze è un passo verso la configurazione in cui la norma degli errori di predizione pesati è minimale.

IV.7 — La pesatura della precisione: il ruolo della matrice di precisione

Un elemento di cruciale importanza filosofica nel framework, che le presentazioni introduttive spesso trascurano, è il ruolo della

matrice di precisione $\Pi^{(l)}$. La precisione è l'inverso della varianza — o, più in generale, l'inverso della matrice di covarianza del rumore — e misura quanto ci si fida del segnale al livello l : un'alta precisione equivale a bassa varianza del rumore, e dunque a un segnale più affidabile. Nel framework bayesiano, la precisione è la misura della certezza con cui si ritiene che l'errore di predizione al livello l sia informativo riguardo agli stati causali sottostanti.

Il termine chiave nelle equazioni di aggiornamento è l'errore di predizione *pesato per la precisione*, spesso denotato $\xi^{(l)}$:

$$\xi^{(l)} = \Pi^{(l)} \epsilon_e^{(l)}$$

Questo errore pesato è il segnale che effettivamente propaga l'informazione verso il livello superiore: non è il semplice errore $\epsilon_e^{(l)}$ ma l'errore moltiplicato per la stima della sua affidabilità. Un errore grande ma con bassa precisione (alta varianza del rumore) produce un segnale ascendente debole; un errore piccolo ma con alta precisione produce un segnale ascendente forte. Le equazioni di aggiornamento diventano:

$$\mu^{(l)} = (\partial \mathbf{g}^{(l)} / \partial \mu^{(l)})^T \xi^{(l)} - \xi^{(l+1)}$$

Questa forma rivela che l'aggiornamento delle credenze al livello l è determinato dalla differenza tra il segnale ascendente pesato dal livello l — che tende ad aggiornare le credenze nella direzione che riduce l'errore di predizione al livello l — e il segnale discendente pesato dal livello $l+1$ — che tende ad ancorare le credenze al livello l alle predizioni del livello superiore.

La matrice di precisione $\Pi^{(l)}$ non è fissa ma può essere essa stessa oggetto di ottimizzazione variazionale. La minimizzazione di $\mathbf{F}$ rispetto a $\Pi^{(l)}$ produce:

$$\Pi^{(l)}_{\text{ottimale}} = (E_Q[\epsilon_e^{(l)} (\epsilon_e^{(l)})^T])^{-1}$$

— la precisione ottimale è l'inverso della covarianza empirica degli errori di predizione. Questo significa che il sistema aggiorna continuamente la propria stima dell'affidabilità dei segnali a ciascun livello in risposta alla distribuzione degli errori di predizione. Un livello che produce errori sistematici e di grande entità riceverà una precisione stimata bassa; un livello che produce errori piccoli e non sistematici riceverà una precisione stimata alta. L'ottimizzazione della precisione è il meccanismo che la codifica predittiva identifica con i processi attentivi: dirigere l'attenzione verso un livello equivale ad aumentare la precisione stimata dei segnali di errore provenienti da quel livello, amplificando la loro influenza sull'aggiornamento delle credenze.

La modulazione della precisione da parte dell'attenzione apre una prospettiva nuova sul problema della coscienza di accesso. Un contenuto diventa cosciente — accessibile ai processi di ragionamento, reportabilità e controllo comportamentale — nella misura in cui riceve alta precisione: nella misura in cui i suoi segnali

di errore influenzano in modo robusto l'aggiornamento delle credenze a tutti i livelli rilevanti. La coscienza di accesso, in questa prospettiva, non è un fenomeno tutto-o-niente ma un continuum determinato dalla distribuzione della precisione attraverso i livelli gerarchici: un contenuto è più o meno cosciente nella misura in cui riceve più o meno precisione, e dunque nella misura in cui influenza in modo più o meno robusto il processo di inferenza. Questo punto sarà sviluppato nel Capitolo VIII in connessione con la teoria dello spazio di lavoro globale di Dehaene e Changeux.

IV.8 — Il filtro di Kalman come caso speciale: connessioni con la teoria del segnale

Per i modelli lineari-gaussiani — in cui le funzioni generative $\mathbf{g}^{(l)}$ sono lineari e le distribuzioni del rumore sono gaussiane — le equazioni della codifica predittiva si riducono al filtro di Kalman, che è il metodo ottimale per l'inferenza su sistemi dinamici lineari-gaussiani. Questa connessione è importante perché il filtro di Kalman ha una lunga storia nella teoria del segnale e nell'ingegneria del controllo, e la sua relazione con la codifica predittiva chiarisce il fondamento matematico del framework.

Nel caso lineare-gaussiano, la funzione generativa è $\mathbf{g}^{(l)}(\mathbf{s}^{(l)}) = \mathbf{A}^{(l)} \mathbf{s}^{(l)}$, dove $\mathbf{A}^{(l)}$ è una matrice di mapping lineare. L'equazione di aggiornamento al livello l è:

$$\boldsymbol{\mu}^{(l)} = (\mathbf{A}^{(l)})^T \boldsymbol{\Pi}^{(l)} (\boldsymbol{\mu}^{(l-1)} - \mathbf{A}^{(l)} \boldsymbol{\mu}^{(l)}) - \boldsymbol{\Pi}^{(l+1)} (\boldsymbol{\mu}^{(l)} - \mathbf{A}^{(l+1)} \boldsymbol{\mu}^{(l+1)})$$

Questa è esattamente l'equazione di aggiornamento del filtro di Kalman per un sistema lineare a due livelli, con il guadagno di Kalman $\mathbf{K}^{(l)} = (\mathbf{A}^{(l)})^T \boldsymbol{\Pi}^{(l)}$ che determina il peso dell'innovazione (l'errore di predizione) nell'aggiornamento dello stato stimato.

Il filtro di Kalman è ottimale nel senso bayesiano per i modelli lineari-gaussiani: produce la vera distribuzione a posteriori, non un'approssimazione variazionale. Questo significa che per i modelli lineari, la codifica predittiva con aggiornamento di gradiente-discesa converge alla soluzione ottimale esatta. Per i modelli non-lineari — che sono quelli di interesse per la cognizione reale — la codifica predittiva è un'approssimazione il cui grado di accuratezza dipende dalla curvatura delle funzioni generative e dalla distanza della distribuzione vera dalla gaussiana.

Questa connessione con il filtro di Kalman ha implicazioni pratiche importanti per le neuroscienze cognitive. Il filtro di Kalman è stato estensivamente studiato in ingegneria del controllo e la sua implementazione neurale è stata proposta da diversi autori come modello dell'elaborazione sensorimotoria. La codifica predittiva generalizza il filtro di Kalman ai modelli non-lineari e alla struttura gerarchica multipla, fornendo un framework unificato che abbraccia tutto lo spettro dall'elaborazione sensoriale elementare alla cognizione astratta.

IV.9 — Il modello generativo come struttura di contenuto: dalla computazione alla semantica

La derivazione matematica del framework predittivo consente ora di enunciare con precisione la tesi filosofica centrale che sarà sviluppata nei capitoli successivi. Il modello generativo gerarchico — la gerarchia di funzioni $\mathbf{g}^{(l)}$, di matrici di precisione $\Pi^{(l)}$, e di distribuzioni a priori $P(\mathbf{s}^{(L)})$ — non è una struttura causale inerte ma una struttura semantica: codifica ipotesi sulle strutture causali dell'ambiente che generano le osservazioni sensoriali. Ogni livello della gerarchia rappresenta una classe di strutture causali a una certa scala di astrazione: i livelli bassi rappresentano proprietà fisiche prossimali dell'input sensoriale (orientazione dei bordi, frequenze sonore, pattern tattili), i livelli intermedi rappresentano oggetti, superfici e agenti, i livelli alti rappresentano relazioni, categorie astratte e strutture narrative.

Il contenuto di una rappresentazione al livello l — la risposta alla domanda «di cosa è questa rappresentazione?» — è determinato da tre elementi congiuntamente:

- (i) la distribuzione a posteriori $Q^{(l)}(\mathbf{s}^{(l)}; \boldsymbol{\mu}^{(l)})$ che il sistema mantiene sugli stati causali al livello l , e che codifica la stima attuale delle cause degli input sensoriali a quella scala;
- (ii) la funzione generativa $\mathbf{g}^{(l)}$ che mappa gli stati al livello l nelle predizioni per il livello $l-1$, e che codifica le assunzioni del sistema sulla struttura causale che connette le entità rappresentate al livello l alle entità rappresentate al livello $l-1$;
- (iii) la matrice di precisione $\Pi^{(l)}$, che determina il peso relativo degli errori di predizione al livello l nell'aggiornamento delle credenze, e che corrisponde all'importanza che il sistema attribuisce alle informazioni provenienti da quel livello.

Questa triplice struttura — distribuzione a posteriori, funzione generativa, matrice di precisione — costituisce ciò che nel seguito chiameremo la *struttura triadica del contenuto predittivo*. Nella prossima sezione si argomenterà che questa struttura risponde ai problemi lasciati aperti dalle teorie di Millikan e Dretske in modo che non è semplicemente un supplemento alle teorie esistenti ma richiede una riconsiderazione del problema del contenuto semantico nella sua interezza.

IV.10 — La normatività endogena: l'errore di predizione come misura intrinseca di correttezza

La tesi più originale che emerge dal framework è quella riguardante la normatività. Come si è argomentato nel Capitolo I, qualsiasi teoria naturalistica del contenuto deve rispondere alla domanda: in virtù di che cosa uno stato rappresentativo può essere *corretto* o

scorretto? La teleosemantica di Millikan risponde attraverso la storia selettiva; la semantica informazionale di Dretske risponde attraverso le dipendenze probabilistiche nei canali normali; il framework predittivo risponde attraverso la struttura interna del modello generativo.

La correttezza di una rappresentazione al livello l , nella prospettiva predittiva, è determinata endogenamente dalla seguente condizione: la rappresentazione $\mu^{(l)}$ è corretta riguardo allo stato causale effettivo $s^{(l)}$ se e solo se il modello generativo, istanziato con $\mu^{(l)}$, produce predizioni per il livello $l-1$ che minimizzano l'errore di predizione pesato $\xi^{(l)} = \Pi^{(l)} \varepsilon_q^{(l)}$. In simboli, la condizione di correttezza per la rappresentazione al livello l è:

$$\|\xi^{(l)}\|^2 = \|\Pi^{(l)}(\mu^{(l-1)} - g^{(l)}(\mu^{(l)}))\|^2 \rightarrow \text{minimo}$$

La normatività non è importata da una storia selettiva esterna né da un canale di informazione ideale: è interna alla struttura del modello generativo. Il sistema *sa* quando le sue rappresentazioni sono scorrette perché produce errori di predizione che non può minimizzare — segnali ascendenti persistenti che indicano che le ipotesi causali del sistema non si accordano con le osservazioni. La correttezza è la configurazione di minima energia libera; la scorrettezza è la deviazione da quella configurazione, misurata dall'energia libera stessa.

Questa internalizzazione della normatività ha una conseguenza filosofica profonda: il problema della misrappresentazione, che per le teorie precedenti richiedeva risorse normative esterne, diventa risolubile all'interno della struttura del sistema. Un sistema predittivo rappresenta scorrettamente $s^{(l)}$ quando la sua distribuzione a posteriori $Q^{(l)}$ è centrata in una regione dello spazio degli stati che produce predizioni sistematicamente discordanti con le osservazioni al livello $l-1$. La scorrettezza è, in questo senso, un fenomeno computazionale osservabile: si manifesta come persistenza degli errori di predizione, come incapacità del sistema di ridurre F al di sotto di una certa soglia. Non richiede di sapere, dall'esterno, qual è il *vero* stato causale: il sistema *sente* la propria scorrettezza attraverso i segnali d'errore non soppressi.

Questa prospettiva consente anche di ridefinire il problema della misrappresentazione con una precisione che le trattazioni precedenti non raggiungevano. La misrappresentazione di tipo normale — percepire un gatto tassidermizzato come un gatto vivo — non è un fallimento del sistema ma un comportamento ottimale dato il modello generativo: il sistema produce le rappresentazioni che minimizzano l'energia libera data la propria struttura generativa e le osservazioni disponibili. La *limitatezza* della misrappresentazione — il fatto che emerga in determinate condizioni e si risolva quando cambiano le informazioni disponibili — è spiegabile in termini di modificazione del modello generativo: nel momento in cui informazioni tattili o olfattive integrano quelle visive, il modello aggiornato produce

rappresentazioni diverse, e l'errore di predizione iniziale viene risolto. La misrappresentazione non è un difetto intrinseco del sistema ma una proprietà emergente dell'interazione tra il modello generativo e l'ambiente, in condizioni in cui le informazioni disponibili sono compatibili con più di una struttura causale.

V — La struttura triadica del contenuto predittivo: dissoluzione del problema della disgiuntività e confronto con Fodor

V.1 — Il problema della disgiuntività riformulato nel quadro predittivo

Il problema della disgiuntività, nella formulazione che ne ha dato Fodor e che si è esaminata nel capitolo precedente, nasce da un'assunzione implicita che le teorie del contenuto ereditate dalla tradizione non hanno mai messo esplicitamente in discussione: l'assunzione che il contenuto di uno stato rappresentativo sia determinato da ciò che *causa* quello stato. Sia nella semantica informazionale di Dretske sia nelle varianti causali più ingenui, il punto di partenza è sempre la relazione che corre tra gli stati del sistema e gli eventi del mondo che li producono. Il problema della disgiuntività è, in questa prospettiva, inevitabile: se il contenuto è determinato dalla relazione causale, e se uno stesso stato può essere prodotto da cause eterogenee, il contenuto dovrà essere la disgiuntiva di tutte quelle cause. Non c'è modo di asimmetrizzare la relazione causale dall'interno della teoria senza introdurre risorse normative esterne.

La proposta che si sviluppa in questo capitolo parte da una riconfigurazione del problema stesso. Nelle reti predittive gerarchiche, la relazione fondamentale non è quella che corre tra gli stati del sistema e le loro cause, ma quella che corre tra gli stati del sistema e le *predizioni* che essi generano. Un sistema predittivo non è primariamente un sistema che risponde a cause: è un sistema che genera ipotesi causali — modelli delle strutture latenti del mondo — e le confronta con le osservazioni disponibili. La direzione rilevante per la determinazione del contenuto non è ascendente, dalla causa allo stato, ma discendente, dallo stato alla predizione. E questa inversione di direzione dissolve il problema della disgiuntività in modo che non è un artificio retorico ma una conseguenza strutturale del framework.

Per rendere questa dissoluzione esplicita, è opportuno partire dall'analisi di un caso concreto — lo stesso tipo di caso che Fodor usava per formulare l'obiezione — e mostrare come si riconfiguri all'interno del framework predittivo. Supponiamo che uno stato rappresentativo $\mu^{(1)}$ al primo livello della gerarchia possa essere prodotto, nel senso dell'indurre una minimizzazione dell'energia

libera in quella direzione, sia dalla presenza di un gatto reale C che da quella di un gatto tassidermizzato T che da quella di un oggetto con sagoma felina S in condizioni di scarsa illuminazione. Nel quadro causale standard, il contenuto di $\mu^{(1)}$ sarebbe $C \vee T \vee S$. Ma nella prospettiva predittiva, la domanda rilevante non è «che cosa causa $\mu^{(1)}$?» bensì «che cosa predice $\mu^{(1)}$?».

La rappresentazione $\mu^{(1)}$, una volta istanziata nel modello generativo, genera una distribuzione di predizioni per il livello sensoriale. Sia $\hat{o} = \mathbf{g}^{(1)}(\mu^{(1)})$ la predizione generata. Questa predizione è specifica: non è una predizione disgiuntiva della forma «ci sarà stimolazione sensoriale coerente con C oppure con T oppure con S », ma una predizione determinata di un pattern sensoriale coerente con le assunzioni del modello su ciò che caratterizza le strutture causali rappresentate da $\mu^{(1)}$. Il contenuto di $\mu^{(1)}$ — nel senso semantico rilevante, ciò di cui la rappresentazione è — è determinato dalle condizioni che renderebbero le predizioni generate da $\mu^{(1)}$ accurate: quelle condizioni, e non l'insieme disgiuntivo delle sue cause passate.

V.2 — La struttura triadica: distribuzione a posteriori, funzione generativa, precisione

Prima di procedere alla dissoluzione formale del problema della disgiuntività, conviene enunciare con precisione la struttura triadica del contenuto predittivo introdotta alla fine del capitolo precedente, perché essa costituisce il cuore della proposta teorica originale di questa monografia.

Si è sostenuto che il contenuto di una rappresentazione al livello l è determinato congiuntamente da tre elementi: la distribuzione a posteriori $Q^{(l)}(\mathbf{s}^{(l)}; \mu^{(l)})$, la funzione generativa $\mathbf{g}^{(l)}$, e la matrice di precisione $\Pi^{(l)}$. Questo non significa semplicemente che il contenuto dipende da tre parametri tecnici del framework: significa che la nozione stessa di contenuto semantico si articola in tre dimensioni distinte che le teorie precedenti non avevano separato con sufficiente rigore.

La prima dimensione è quella che potremmo chiamare il *contenuto proposizionale*: ciò che la rappresentazione afferma riguardo agli stati causali del mondo al livello di astrazione corrispondente. Questa dimensione è codificata dalla distribuzione a posteriori $Q^{(l)}$: il parametro $\mu^{(l)}$ che indica la stima a posteriori degli stati causali determina *quale* struttura causale il sistema sta rappresentando. Se $\mu^{(1)}$ è centrato sulla rappresentazione di un gatto adulto in posizione eretta, il contenuto proposizionale della rappresentazione è approssimativamente «c'è un gatto adulto in posizione eretta».

La seconda dimensione è quella che chiameremo il *contenuto inferenziale*: l'insieme delle relazioni di implicazione e di incompatibilità che la rappresentazione intrattiene con altre rappresentazioni, determinato dalla funzione generativa $\mathbf{g}^{(l)}$. La

funzione generativa codifica le assunzioni del sistema sulla struttura causale che connette il livello l al livello $l-1$: codifica, in altri termini, le *conseguenze* della rappresentazione per i livelli inferiori della gerarchia. Il contenuto inferenziale di «c'è un gatto adulto in posizione eretta» include le implicazioni che derivano dalle assunzioni generative del sistema: ci sarà una certa distribuzione di profili, di texture, di pattern di movimento, di suoni, che il modello predice coerentemente con quella struttura causale.

La terza dimensione è quella del *contenuto attentivo*: il peso relativo che la rappresentazione riceve nell'elaborazione complessiva del sistema, determinato dalla matrice di precisione $\Pi^{(l)}$. Questa dimensione non ha un corrispettivo nelle teorie tradizionali del contenuto, che tendono a trattare il contenuto come una proprietà binaria — uno stato o ha un certo contenuto o non lo ha — indipendentemente dall'importanza che quella rappresentazione riveste nel sistema complessivo. La precisione introduce invece una dimensione quantitativa: la stessa rappresentazione proposizionale può avere un peso attentivo diverso a seconda del contesto, e questa variazione è semanticamente rilevante nel senso che determina quale parte dell'input sensoriale verrà elaborata in modo privilegiato e quale sarà soppressa o relegata a livelli meno influenti della gerarchia.

La distinzione tra le tre dimensioni del contenuto ha implicazioni dirette per alcune questioni classiche della filosofia del linguaggio e della mente. La distinzione tra contenuto proposizionale e contenuto inferenziale è analoga, ma non identica, alla distinzione fregeana tra riferimento e senso. Il contenuto proposizionale corrisponde approssimativamente al riferimento: è la struttura causale nel mondo che la rappresentazione coglie. Il contenuto inferenziale corrisponde approssimativamente al senso: è la rete di implicazioni e aspettative attraverso cui il sistema si rapporta al riferimento. La differenza cruciale è che nel framework predittivo entrambe le dimensioni sono specificate in termini computazionali precisi — la distribuzione a posteriori e la funzione generativa — invece di rimanere a livello di descrizione qualitativa.

Il contenuto attentivo, d'altro canto, non ha un corrispettivo diretto nella tradizione fregeana o nella semantica intenzionale. È una dimensione genuinamente nuova che emerge dalla struttura della codifica predittiva e che riflette il fatto che il contenuto non è una proprietà statica delle rappresentazioni ma una proprietà dinamica che dipende dall'importanza relativa che le diverse rappresentazioni ricevono nel processo computazionale. In termini fenomenologici, questa dimensione corrisponde alla distinzione tra ciò che è al fuoco dell'attenzione e ciò che è sullo sfondo: lo stesso oggetto ha un contenuto diverso quando è al centro dell'attenzione rispetto a quando è percepito marginalmente, non perché cambino le sue proprietà ma perché cambia il peso che le sue caratteristiche ricevono nell'elaborazione complessiva.

V.3 — Dissoluzione formale del problema della disgiuntività

La dissoluzione del problema della disgiuntività che il framework predittivo consente può essere esposta con precisione attraverso il seguente argomento formale. Si consideri uno stato rappresentativo $\mu^{(1)}$ che sia causato, nelle diverse occasioni di occorrenza, sia dal tipo di stimolo C (gatti reali) sia dal tipo di stimolo T (gatti tassidermizzati) sia dal tipo di stimolo S (sagome feline in scarsa luce). La teoria causale del contenuto asserisce che il contenuto di $\mu^{(1)}$ è la disgiuntiva $C \vee T \vee S$, poiché tutti e tre i tipi di stimolo producono quello stato.

Nel framework predittivo, il contenuto di $\mu^{(1)}$ è determinato dalle predizioni che esso genera attraverso la funzione generativa $g^{(1)}$. Siano $\mathbf{o}_C$, $\mathbf{o}_T$, $\mathbf{o}_S$ le distribuzioni di osservazioni sensoriali associate rispettivamente ai tipi C, T, S. Le tre distribuzioni sono *distinte*: un gatto vivo, un gatto tassidermizzato e una sagoma felina producono pattern sensoriali differenti, anche se sufficientemente simili da indurre stati $\mu^{(1)}$ simili (o identici, nel caso limite) al livello gerarchico considerato. Ora, la funzione generativa $g^{(1)}(\mu^{(1)})$ produce una predizione determinata $\hat{\mathbf{o}} = g^{(1)}(\mu^{(1)})$, non una distribuzione disgiuntiva. Questa predizione si troverà in una relazione diversa con le tre distribuzioni osservative $\mathbf{o}_C$, $\mathbf{o}_T$, $\mathbf{o}_S$: la distanza

$$d(\hat{\mathbf{o}}, \mathbf{o}_C) = \|\Pi^{(0)}(\mathbf{o}_C - g^{(1)}(\mu^{(1)}))\|^2$$

non è in generale uguale a $d(\hat{\mathbf{o}}, \mathbf{o}_T)$ né a $d(\hat{\mathbf{o}}, \mathbf{o}_S)$.

Se il modello generativo è stato costruito — attraverso l'apprendimento o la selezione — in un ambiente in cui C era il tipo di stimolo prevalente che produceva $\mu^{(1)}$, allora $g^{(1)}$ sarà calibrata in modo tale che $g^{(1)}(\mu^{(1)}) \approx \mathbf{o}_C$: la predizione generata da $\mu^{(1)}$ si accorda meglio con le osservazioni associate ai gatti reali che con quelle associate ai gatti tassidermizzati o alle sagome. In questo senso, il modello generativo introduce un'asimmetria strutturale tra le diverse cause di $\mu^{(1)}$: C è la causa *canonica* nel senso che il modello generativo è calibrato su C, mentre T e S sono cause *atipiche* nel senso che producono $\mu^{(1)}$ ma si discostano dalla condizione che il modello predice.

L'asimmetria tra causa canonica e cause atipiche si riflette in modo quantitativo nell'errore di predizione che emerge nelle diverse condizioni. Quando la causa effettiva è C, l'errore di predizione al livello sensoriale — $\varepsilon^{(0)} = \mathbf{o} - g^{(1)}(\mu^{(1)})$ — è piccolo e non sistematico: il modello predice correttamente le osservazioni sensoriali associate al gatto reale. Quando la causa effettiva è T, l'errore di predizione è sistematicamente più elevato nelle dimensioni sensoriali che distinguono un gatto tassidermizzato da uno vivo — l'assenza di movimento respiratorio, la diversa risposta termica, la diversa risposta alle stimolazioni tattili — ed è questo errore sistematico persistente che indica al sistema che la sua

rappresentazione è parzialmente inadeguata. Non basta avere una rappresentazione che minimizza l'energia libera al livello visivo; la rappresentazione deve essere valutata rispetto all'intero profilo di predizioni multimodali che essa genera.

La dissoluzione formale si completa osservando che il contenuto di $\mu^{(1)}$ — determinato dalle condizioni che le predizioni di $\mu^{(1)}$ rendono più probabili — è necessariamente non-disgiuntivo: le predizioni di $\mu^{(1)}$ sono specifiche, non disgiuntive, e le condizioni che le rendono accurate sono specifiche, non disgiuntive. La disgiuntività dell'insieme delle cause è un fatto sul processo causale ascendente, non sul contenuto della rappresentazione; il contenuto è determinato dal processo generativo discendente, che è non-disgiuntivo per costruzione.

È importante sottolineare che questa dissoluzione non è semplicemente una ridefinizione del termine «contenuto» per aggirare il problema. La scelta di determinare il contenuto attraverso le predizioni piuttosto che attraverso le cause è motivata da argomenti indipendenti: le predizioni catturano ciò che la rappresentazione «significa» per il sistema nel senso che determina come il sistema si comporterà in conseguenza di essa — quali aspettative avrà, come risponderà a stimoli futuri, come integrerà la rappresentazione con altre credenze. Le cause di uno stato non determinano il suo significato nel senso rilevante: un'allucinazione può avere le stesse cause proximal di una percezione normale ma avere un significato diverso — un significato errato. Le predizioni, al contrario, determinano il significato nel senso rilevante: un'allucinazione ha le stesse predizioni di una percezione veridica e si distingue da essa solo per il fatto che le sue predizioni non sono soddisfatte dalle osservazioni successive.

V.4 — Il ruolo della gerarchia nella determinazione del contenuto

La dissoluzione del problema della disgiuntività appena esposta funziona in modo particolarmente elegante se si tiene conto della struttura gerarchica del modello generativo, che è l'aspetto cruciale rispetto alle teorie monolevelliche del contenuto. Nelle teorie di Dretske e di Millikan, il contenuto di una rappresentazione è determinato dalla sua relazione con gli eventi del mondo a un determinato livello di analisi. Nella codifica predittiva gerarchica, ogni rappresentazione al livello 1 è in relazione non solo con le osservazioni sensoriali al livello 0 — la relazione terminale della catena predittiva — ma anche con le rappresentazioni a tutti i livelli inferiori lungo la catena generativa. Questa relazione mediata è essenziale per la determinazione del contenuto.

Consideriamo la struttura gerarchica a tre livelli: al livello 2 si trovano rappresentazioni di categorie di oggetti (gatto, cane, uccello); al livello 1 si trovano rappresentazioni di proprietà morfologiche (forma del corpo, texture del pelo, schema

locomotore); al livello 0 le osservazioni sensoriali. La rappresentazione $\mu^{(2)}$ di «gatto» genera, attraverso $g^{(2)}$, predizioni sulle rappresentazioni al livello 1: il modello prevede che un gatto abbia una certa distribuzione di proprietà morfologiche. Queste predizioni al livello 1 a loro volta generano, attraverso $g^{(1)}$, predizioni sulle osservazioni sensoriali al livello 0.

Il contenuto di $\mu^{(2)}$ — ciò di cui essa è una rappresentazione — non è determinato soltanto dalla relazione tra $\mu^{(2)}$ e le osservazioni sensoriali finali, ma dall'intera catena di predizioni che essa genera attraverso la gerarchia. Questo ha una conseguenza importante per il problema della disgiuntività: due stimoli C e T che producono lo stesso stato $\mu^{(2)}$ al livello 2 potrebbero produrre stati diversi al livello 1, e le predizioni generate da $\mu^{(2)}$ per il livello 1 potrebbero accordarsi meglio con uno dei due stimoli. In questo caso, il modello generativo al livello 2 non è agnostico tra C e T: esso prevede un profilo di proprietà morfologiche che è più simile a quello di C che a quello di T, e questo accordo differenziale con i diversi stimoli è ciò che asimmetrizza la relazione causale e determina il contenuto.

Formalmente, la probabilità congiunta degli stati a tutti i livelli della gerarchia, calcolata con il modello generativo, è:

$$P(o, s^{(1)}, s^{(2)}) = P(o \mid s^{(1)}) \cdot P(s^{(1)} \mid s^{(2)}) \cdot P(s^{(2)})$$

Il contenuto di $\mu^{(2)}$ è determinato dalle condizioni in cui l'aspettativa marginale del modello, calcolata rispetto agli stati al livello 1 e alle osservazioni, è massimizzata:

$$\begin{aligned} s^{(2)*} &= \arg \max_s E[\log P(o, s^{(1)}, s^{(2)} = s) \mid \mu^{(2)}] \\ &= \arg \max_s \{ E[\log P(o \mid s^{(1)})] + E[\log P(s^{(1)} \mid s)] + \log P(s) \} \end{aligned}$$

dove l'aspettativa è presa rispetto alla distribuzione $Q^{(1)}(s^{(1)}; \mu^{(1)})$ indotta dalla rappresentazione corrente al livello 1. Il contenuto di $\mu^{(2)}$ è l'insieme delle condizioni nel mondo che massimizzano questa aspettativa — non l'insieme disgiuntivo delle sue cause passate.

V.5 — Confronto con il programma di Fodor: il linguaggio del pensiero e la tesi della composizionalità

Jerry Fodor rappresenta il punto di riferimento critico principale per qualsiasi proposta che intenda fondare una teoria naturalistica del contenuto semantico nelle strutture computazionali della mente. Il suo programma — articolato nel corso di quattro decenni a partire da *The Language of Thought* (1975) fino a *LOT 2: The Language of Thought Revisited* (2008) — combina una teoria dell'architettura mentale (il linguaggio del pensiero, o LOT) con una teoria del contenuto (la semantica intenzionale asimmetrica, o SIA) e con una teoria della struttura concettuale (l'atomismo concettuale). La critica che la codifica predittiva rivolge a questo

programma non è complessiva ma specifica: mira a mostrare che le assunzioni architetturali del LOT — in particolare la composizionalità del contenuto e il carattere atomistico dei concetti primitivi — sono incompatibili con la struttura del contenuto semantico quale emerge dal framework predittivo.

Il LOT sostiene che i pensieri sono strutture composizionali: un pensiero complesso è composto di pensieri più semplici, e il suo contenuto è determinato dal contenuto delle parti e dalla struttura sintattica che le combina. La composizionalità è per Fodor una proprietà indefettibile della mente cognitiva: la produttività del pensiero — il fatto che siamo in grado di formare e comprendere un numero illimitato di pensieri nuovi — e la sistematicità — il fatto che la capacità di formare il pensiero «A ama B» implica la capacità di formare il pensiero «B ama A» — richiedono entrambe che i pensieri abbiano struttura composizionale. La composizionalità del pensiero richiede a sua volta che i concetti elementari che fungono da componenti abbiano contenuti stabili e indipendenti dal contesto: il concetto GATTO deve avere lo stesso contenuto in «il gatto è sul tappeto» e in «il gatto non è sul tappeto» e in «se il gatto è sul tappeto allora...». I contenuti dei concetti primitivi non sono determinati dal contesto composizionale in cui compaiono, ma sono fissi — atomistici.

La semantica intenzionale asimmetrica, che Fodor ha sviluppato in *A Theory of Content* (1990), tenta di fondare i contenuti atomistici degli input primitivi in relazioni causali asimmetriche: il concetto COW ha MUCCA come contenuto perché la legge causale che connette mucche al tokening di COW è asimmetricamente dipendente e non viceversa — se le oche causassero COW, lo farebbero perché le oche assomigliano alle mucche in modi causalmente rilevanti, mentre non vale il contrario. La SIA è una risposta sofisticata al problema della disgiuntività, ma dipende crucialmente dall'atomismo dei concetti primitivi: le leggi causali asimmetriche determinano il contenuto solo per i concetti che sono causalmente collegati in modo diretto con le loro denotazioni, non per i concetti astratti o relazionali il cui contenuto non è determinato da alcuna legge causale del tipo richiesto.

La critica che il framework predittivo rivolge all'atomismo concettuale di Fodor è di carattere strutturale. Nella codifica predittiva, il contenuto di una rappresentazione a un dato livello gerarchico non è atomistico ma è determinato dalla sua posizione nella rete di relazioni generative con tutti gli altri livelli della gerarchia. Il concetto GATTO non ha un contenuto fisso e indipendente dal contesto: il suo contenuto è in parte determinato dalle relazioni generative che lo connettono verso il basso alle rappresentazioni di proprietà morfologiche, verso l'alto alle rappresentazioni di categorie biologiche, e lateralmente alle rappresentazioni di altri concetti animali. Questa dipendenza contestuale non è una debolezza ma una caratteristica strutturale che risponde a evidenze empiriche sulla variabilità contestuale del

contenuto concettuale — evidenze che l'atomismo fodoriano non riesce a incorporare senza sforzo teorico considerevole.

La composizionalità, d'altra parte, non viene abbandonata ma ricollocata: le reti predittive gerarchiche sono composizionali nel senso che il modello generativo è costruito da componenti che si combinano in modo regolare, e le predizioni generate da rappresentazioni complesse sono determinate in modo sistematico dalle predizioni generate dalle rappresentazioni componenti. La sistematicità e la produttività — le proprietà che Fodor adduce come argomenti per la composizionalità — sono preserved nel framework predittivo, ma attraverso la struttura composizionale del modello generativo piuttosto che attraverso la struttura sintattica delle rappresentazioni nel senso del LOT.

La differenza cruciale tra la composizionalità del LOT e la composizionalità del modello generativo è la seguente. Nel LOT, la composizionalità è sintattica: il contenuto di un pensiero complesso è determinato dal contenuto delle parti e dalla struttura sintattica che le combina, indipendentemente dal contesto computazionale in cui il pensiero viene prodotto. Nella codifica predittiva, la composizionalità è generativa: il contenuto di una rappresentazione complessa è determinato dalle predizioni che essa genera attraverso la gerarchia, e queste predizioni sono composte in modo sistematico dalle predizioni dei componenti — ma il modo in cui le predizioni si compongono dipende dalla struttura del modello generativo, non da regole sintattiche indipendenti dal contenuto. La composizionalità generativa è contestuale nel senso che le predizioni generate da un concetto dipendono dalla struttura dell'intero modello, non solo dal concetto stesso; ma è sistematica nel senso che le relazioni tra le predizioni dei componenti e le predizioni del tutto sono determinate in modo regolare dalla struttura del modello.

VI — Il grounding del contenuto astratto: dalla percezione all'astrazione attraverso la gerarchia generativa

VI.1 — Il problema del grounding e la sfida dell'astratto

Il problema del grounding semantico — come le rappresentazioni simboliche di un sistema cognitivo acquistino connessione con il mondo, come non siano meri simboli non interpretati in un sistema formale chiuso in sé stesso — fu formulato da Harnad nel 1990 nel contesto del dibattito sull'IA classica, ma ha radici più profonde nella questione brentaniana dell'intenzionalità. Per Harnad, il problema del grounding si manifesta in forma acuta nel cosiddetto *symbol grounding problem*: in un sistema di manipolazione

di simboli, i simboli sono interpretati attraverso altri simboli in una catena potenzialmente infinita che non tocca mai la realtà che i simboli dovrebbero denotare. Il dizionario, come argomento paradigmatico, definisce le parole mediante altre parole: il significato di «gatto» è spiegato attraverso «felino», «mammifero», «carnivoro», «domestico», e ciascuno di questi termini è a sua volta spiegato attraverso altri termini. Un sistema cognitivo che non avesse accesso diretto a nessuna delle realtà denotate dai simboli del suo lessico non capirebbe nulla — manipolerebbe simboli non interpretati.

La risposta standard, e in larga misura corretta, al problema del grounding è che i simboli di livello elementare sono ancorati direttamente alla realtà attraverso l'esperienza percettiva: il concetto ROSSO è ancorato all'esperienza visiva del rosso, il concetto CALDO all'esperienza termica, il concetto GATTO alla percezione di gatti. I simboli astratti — NUMERO, GIUSTIZIA, CAUSALITÀ — sono poi costruiti a partire dai simboli elementari attraverso operazioni composizionali e relazionali. Questa risposta, pur plausibile per i concetti elementari, incontra difficoltà crescenti man mano che ci si sposta verso concetti di maggiore astrazione: non è chiaro come CAUSALITÀ possa essere ancorato all'esperienza percettiva in modo che ne determini il contenuto in modo non circolare, né come l'ancoraggio percettivo dei concetti elementari si trasmetta ai concetti astratti attraverso la composizione.

Le difficoltà dell'approccio composizionale al grounding sono state documentate in dettaglio dalla psicologia cognitiva dello sviluppo e dalla neuropsicologia. I bambini non acquisiscono i concetti astratti semplicemente componendo concetti percettivi elementari: attraversano stadi di sviluppo in cui la struttura del concetto astratto si modifica qualitativamente, suggerendo che i concetti astratti hanno una struttura propria che non è riducibile alla semplice composizione dei componenti percettivi. I pazienti con lesioni cerebrali specifiche possono perdere la capacità di elaborare certi concetti astratti mantenendo intatta la capacità di elaborare i concetti percettivi componenti, suggerendo che i concetti astratti sono rappresentati in strutture neurali parzialmente separate e non semplicemente nelle strutture dei loro componenti.

Il framework della codifica predittiva offre una risposta al problema del grounding che è strutturalmente diversa dalla risposta composizionale standard. Il grounding non avviene attraverso una catena dal basso verso l'alto — dai percetti ai concetti — ma attraverso una struttura bidirezionale che percorre la gerarchia in entrambe le direzioni. I concetti astratti ai livelli superiori della gerarchia non derivano il loro contenuto dai percetti ai livelli inferiori attraverso aggregazione composizionale: sono essi stessi parte della struttura generativa che *produce* le predizioni per i livelli inferiori, e il loro contenuto è determinato dalla struttura di quelle predizioni. Il grounding è quindi non un ancoraggio

puntuale dei simboli all'esperienza sensoriale, ma una relazione strutturale pervasiva: ogni livello della gerarchia è ancorato ai livelli inferiori attraverso la funzione generativa, e questa relazione di ancoraggio si estende dall'alto verso il basso dell'intera gerarchia.

VI.2 — La distinzione tra grounding percettivo e grounding strutturale

Per articolare questa proposta con la precisione necessaria, occorre distinguere due tipi di grounding che il framework predittivo mette in gioco e che la discussione tradizionale ha spesso confuso. Il grounding percettivo è la relazione che i livelli bassi della gerarchia intrattengono con l'input sensoriale: le rappresentazioni al livello 1 sono direttamente vincolate dalle osservazioni sensoriali attraverso la funzione di verosimiglianza $P(\mathbf{o} \mid \mathbf{s}^{(1)})$, e il loro contenuto è in parte determinato da questo vincolo. Il grounding strutturale è la relazione che i livelli superiori intrattengono con i livelli inferiori attraverso le funzioni generative $\mathbf{g}^{(l)}$: le rappresentazioni ai livelli superiori sono vincolate dalle rappresentazioni ai livelli inferiori in modo che il loro contenuto è determinato dalla struttura delle trasformazioni generative che li connettono.

Il grounding percettivo e il grounding strutturale hanno proprietà diverse che è importante tenere distinte. Il grounding percettivo è essenzialmente *locale* e *sincrono*: dipende dall'input sensoriale attuale e agisce istantaneamente attraverso l'aggiornamento delle distribuzioni a posteriori. Il grounding strutturale è *globale* e *diacronico*: dipende dalla struttura del modello generativo, che è il prodotto dell'intera storia di interazione del sistema con l'ambiente — attraverso l'apprendimento, l'adattamento, e in ultima istanza la storia evolutiva — e che si modifica lentamente attraverso l'esperienza accumulata.

Questa distinzione consente di rispondere a una obiezione naturale che potrebbe essere rivolta al framework predittivo: se il contenuto dei concetti astratti è determinato dalla struttura generativa, e se la struttura generativa può in principio essere diversa per sistemi diversi, il contenuto dei concetti astratti risulterebbe soggettivo o sistema-relativo in modo inaccettabile. La risposta è che il grounding strutturale, pur variando tra sistemi con storie di apprendimento diverse, è vincolato dal grounding percettivo in modo che produce convergenza sul contenuto per sistemi esposti a ambienti simili. Due sistemi che abbiano appreso il modello generativo attraverso l'interazione con lo stesso tipo di ambiente convergeranno verso strutture generative simili, e dunque verso contenuti simili per i loro concetti astratti, anche se le strutture generative non sono identiche e il contenuto non è rigorosamente lo stesso nei due sistemi.

La convergenza intersoggettiva del contenuto astratto è in realtà un fenomeno empirico ben documentato: i sistemi cognitivi umani, esposti ad ambienti fisici e sociali simili, sviluppano concetti

astratti con strutture sorprendentemente simili. Le evidenze cross-culturali sulla cognizione dei colori, della causalità, del numero, dello spazio mostrano sia universali — aspetti del contenuto concettuale che convergono in tutte le culture — sia variazioni — aspetti che riflettono le specifiche strutture ambientali e sociali in cui il sistema è cresciuto. Il framework predittivo prevede esattamente questa struttura: gli universali riflettono i vincoli imposti dal grounding percettivo comune (tutti gli esseri umani hanno lo stesso sistema sensoriale che li espone alle stesse strutture fisiche del mondo) e le variazioni riflettono le differenze nel grounding strutturale (sistemi cresciuti in ambienti sociali e linguistici diversi costruiscono strutture generative diverse).

VI.3 — La causalità come concetto di livello gerarchico alto: un caso di studio

Per illustrare concretamente come il framework predittivo account per il grounding dei concetti astratti, si esaminerà in dettaglio il caso del concetto CAUSALITÀ, che è particolarmente rilevante dato il tema di questo lavoro e che presenta caratteristiche esemplari per la discussione.

Il concetto di causalità occupa nella gerarchia cognitiva umana un livello di astrazione molto elevato: è un concetto che si applica a relazioni tra eventi di qualsiasi tipo, che è invariante rispetto alle modalità sensoriali, che richiede l'integrazione di informazioni temporali e spaziali provenienti da livelli multipli della gerarchia percettiva. Eppure, come Michotte ha dimostrato sperimentalmente già nel 1946 con i suoi esperimenti sui «lanci» percettivi, gli esseri umani percepiscono la causalità in modo diretto e immediato in certi tipi di sequenze di eventi: quando una forma geometrica si muove verso un'altra e questa si mette in moto al contatto, i soggetti riportano la percezione di una relazione causale, non semplicemente di una sequenza temporale. La percezione della causalità — la fenomenologia della «connessione necessaria» che Hume cercava invano nelle impressioni sensoriali — è un fenomeno cognitivo robusto che ha correlati neurali identificabili.

La questione humiana è importante non solo come problema storico ma come sfida sistematica a qualsiasi teoria del grounding della causalità. Hume sosteneva che nell'esperienza sensoriale non si trova nessuna impressione di «connessione necessaria» tra causa ed effetto: si trova solo la costante congiunzione di eventi dello stesso tipo, la successione regolare, non la necessità. La connessione necessaria che costituisce l'essenza della relazione causale è una proiezione della mente sull'esperienza, non una caratteristica dell'esperienza stessa. Questa posizione ha generato la domanda: dove si trova la necessità causale, se non nell'esperienza? Come si gronda il concetto di causalità, se l'esperienza non mostra la necessità che il concetto richiede?

Come si inserisce questo fenomeno nella struttura della codifica predittiva gerarchica? La proposta è la seguente. Il concetto CAUSALITÀ corrisponde a una famiglia di rappresentazioni ai livelli superiori della gerarchia — diciamo, ai livelli 3 o 4 in una gerarchia a sei livelli — il cui contenuto è determinato dalla struttura delle predizioni che generano per i livelli inferiori. Queste predizioni riguardano le *relazioni temporali e spaziali* tra eventi ai livelli inferiori: in particolare, predicono che se un evento E_1 di un certo tipo è rappresentato al livello 2, allora successivamente e con alta probabilità un evento E_2 di un certo tipo correlato sarà rappresentato al livello 2, e questa covariazione sarà sistematica e generalizzabile a nuovi contesti.

Formalmente, il contenuto del concetto CAUSALITÀ(E_1, E_2) al livello n può essere espresso come la condizione sulla struttura del modello generativo che rende alta la probabilità condizionale:

$$P(s^{(2)}(t+\delta t) \in R_{\{E_2\}} \mid s^{(2)}(t) \in R_{\{E_1\}}, s^{(n)} = \text{CAUSALITÀ}) \\ \gg P(s^{(2)}(t+\delta t) \in R_{\{E_2\}})$$

dove $R_{\{E_1\}}$ e $R_{\{E_2\}}$ sono regioni dello spazio degli stati al livello 2 corrispondenti ai tipi di eventi E_1 ed E_2 , e δt è l'intervallo temporale appropriato. La rappresentazione CAUSALITÀ(E_1, E_2) è la struttura di livello n che genera predizioni per il livello 2 tali da rendere altamente probabile la sequenza $E_1 \rightarrow E_2$. Quando questa predizione è soddisfatta — quando la sequenza osservata corrisponde alla predizione — l'errore di predizione è basso e la rappresentazione causale è confermata; quando la predizione non è soddisfatta, l'errore di predizione è alto e la rappresentazione causale deve essere aggiornata o abbandonata.

Questo account del concetto CAUSALITÀ ha una proprietà notevole: esso risponde, in termini computazionali, alla domanda che Hume poneva in termini empiristi. Hume non trovava la «connessione necessaria» nell'impressione diretta perché cercava una qualità percettiva localizzata in un singolo evento; nel framework predittivo, la rappresentazione causale non è localizzata in nessun singolo evento ma è una struttura di livello superiore che genera predizioni sulle *relazioni temporali* tra eventi. La «necessità» della connessione causale non è una qualità percettiva ma una caratteristica del modello generativo: è la struttura che rende alta la probabilità predittiva della sequenza $E_1 \rightarrow E_2$. Quando la struttura è robusta — quando il modello ha appreso, attraverso molte istanze, che E_1 è seguito da E_2 in modo sistematico — la predizione della sequenza è «necessaria» nel senso che deviazioni da essa producono errori di predizione elevati, che il sistema è motivato a ridurre.

La connessione tra la risposta predittiva alla domanda humiana e la fenomenologia michottiana della percezione diretta della causalità è la seguente. Michotte osservava che la causalità è percepita in modo diretto — non inferita riflessivamente — in certe configurazioni di stimoli. Nel framework predittivo, questa

percezione diretta corrisponde all'attivazione di rappresentazioni causali ai livelli superiori della gerarchia in risposta a pattern di movimento che soddisfano le predizioni di quei livelli: il sistema non deve «ragionare» sulla causalità perché ha già un modello generativo che prevede le relazioni causali e le rileva automaticamente attraverso il processo di minimizzazione dell'energia libera. La percezione diretta della causalità è il fenomeno soggettivo corrispondente all'abbassamento dell'errore di predizione causale in risposta a una sequenza di eventi causalmente congruente.

VI.4 — L'apprendimento del modello generativo: aggiornamento sinaptico e metaplasticità

Il framework predittivo fin qui presentato descrive l'inferenza percettiva su tempi brevi — l'aggiornamento delle stime a posteriori $\mu^{(i)}$ in risposta agli input sensoriali, dati i parametri del modello generativo. Ma il modello generativo stesso — le funzioni $g^{(i)}$ e le matrici di precisione $\Pi^{(i)}$ — non è fisso: è il prodotto di un processo di apprendimento che si svolge su tempi più lunghi. Comprendere la dinamica dell'apprendimento è essenziale per comprendere come il contenuto dei concetti astratti si formi e si modifichi nel corso dell'esperienza.

L'apprendimento nel framework della codifica predittiva può essere visto come minimizzazione dell'energia libera rispetto ai parametri del modello generativo, piuttosto che rispetto agli stati rappresentazionali. Mentre l'inferenza percettiva minimizza F rispetto a $\mu^{(i)}$ tenendo fissi i parametri del modello, l'apprendimento minimizza F rispetto ai parametri θ del modello — i pesi sinaptici che implementano le funzioni $g^{(i)}$ — tenendo fisse le stime a posteriori:

$$\dot{\theta} = -\eta \partial F / \partial \theta$$

dove η è il tasso di apprendimento. Il gradiente dell'energia libera rispetto ai parametri del modello è:

$$\partial F / \partial \theta = -\sum (\partial g^{(i)}(\mu^{(i)}; \theta) / \partial \theta)^T \Pi^{(i)} \epsilon_q^{(i)}$$

Questa regola di apprendimento è una forma di apprendimento hebbiano pesato per la precisione: i parametri del modello sono aggiornati in direzione che riduce gli errori di predizione pesati, proporzionalmente alla sensibilità della funzione generativa rispetto ai parametri stessi. La struttura matematica di questa regola è analoga alla retropropagazione del gradiente nelle reti neurali artificiali, ma con una differenza cruciale: il segnale d'errore che guida l'aggiornamento non è un segnale supervisorio esterno ma è l'errore di predizione interno del sistema — la discrepanza tra le predizioni top-down e le osservazioni bottom-up. Il sistema apprende minimizzando la propria sorpresa, non ottimizzando una funzione obiettivo esternamente specificata. Questo carattere non-supervisivo dell'apprendimento è una delle proprietà

biologicamente più plausibili del framework: il cervello non dispone di un «insegnante esterno» che fornisca segnali d'errore supervisor, e deve imparare esclusivamente dalla struttura dei propri input sensoriali.

La metaplasticità — la regolazione dei parametri della plasticità sinaptica da parte dell'esperienza passata — corrisponde nel framework all'aggiornamento delle matrici di precisione $\Pi^{(t)}$. Il sistema non apprende solo quali strutture causali esistono nell'ambiente (aggiornamento di θ) ma anche quanto ci si può fidare dei segnali a ciascun livello gerarchico (aggiornamento di Π). Questa doppia modalità di apprendimento — dell'errore e della precisione — costituisce la base computazionale per fenomeni come l'adattamento sensoriale, l'effetto di ordine della presentazione degli stimoli, e la modulazione dell'apprendimento da parte dello stato emotivo.

Le regole di apprendimento per la precisione possono essere derivate analogamente alla regola per i parametri del modello. Minimizzando $\mathbf{F}$ rispetto a $\Pi^{(t)}$:

$$\partial \mathbf{F} / \partial \Pi^{(t)} = \frac{1}{2} (\mathbf{E}_Q[\epsilon_e^{(t)} (\epsilon_e^{(t)})^T] - (\Pi^{(t)})^{-1}) = 0$$

che produce la condizione di stazionarietà:

$$\Pi^{(t)}_{\text{ottimale}} = (\mathbf{E}_Q[\epsilon_e^{(t)} (\epsilon_e^{(t)})^T])^{-1}$$

La precisione ottimale è l'inverso della covarianza media degli errori di predizione. Questa regola aggiorna la precisione in modo da riflettere la distribuzione empirica degli errori: livelli con errori grandi e sistematici ricevono bassa precisione, livelli con errori piccoli e non sistematici ricevono alta precisione. L'aggiornamento della precisione è graduale — su tempi più lunghi dell'aggiornamento dei parametri del modello — poiché stima una statistica della distribuzione degli errori che richiede molte osservazioni per essere stabile.

VI.5 — GIUSTIZIA come concetto valutativo: grounding normativo e varianza contestuale

Il secondo caso di studio per l'analisi del grounding strutturale riguarda il concetto GIUSTIZIA, che appartiene a una classe radicalmente diversa tanto da CAUSALITÀ quanto da NUMERO: è un concetto valutativo, connesso a pratiche sociali, a sistemi normativi e a intuizioni morali che variano significativamente tra individui, culture e contesti storici. La domanda se concetti valutativi come GIUSTIZIA abbiano contenuto determinato nel senso della teleosemantica computazionale sincronica, o se la loro variabilità contestuale non li renda inesorabilmente indeterminati o relativi, è una domanda che il framework deve affrontare.

Gli studi di psicologia morale di Haidt e Greene hanno mostrato che i giudizi morali umani — inclusi i giudizi di giustizia e ingiustizia — hanno una struttura doppia: una componente

intuitiva, rapida, emotivamente connotata, che precede e spesso determina la valutazione conscia; e una componente razionale, lenta, riflessiva, che rielabora e talvolta corregge l'intuizione iniziale. L'evidenza neuroimaging mostra che la valutazione di scenari di ingiustizia attiva sistematicamente aree associate all'elaborazione emotiva — amigdala, insula anteriore, corteccia orbitofrontale — oltre alle aree prefrontali associate al ragionamento normativo.

Nel framework predittivo, questa struttura doppia corrisponde a due livelli gerarchici distinti che contribuiscono al giudizio di giustizia. Al livello della valutazione intuitiva — corrispondente ai livelli medio-bassi della gerarchia — vi sono rappresentazioni di distribuzioni di risorse, di trattamenti ricevuti e di aspettative violate che attivano segnali d'errore affettivamente connotati: la percezione di una distribuzione ineguale produce un errore di predizione in un sistema che ha appreso, attraverso l'esperienza sociale, che certe distribuzioni violano aspettative radicate. Al livello della valutazione riflessiva — corrispondente ai livelli superiori della gerarchia — vi sono rappresentazioni di principi normativi, di teorie della giustizia, di sistemi di norme socialmente condivise, che generano predizioni sulle distribuzioni appropriate e producono segnali d'errore cognitivi quando quelle distribuzioni vengono violate.

La variabilità culturale dei giudizi di giustizia corrisponde, in questa prospettiva, alla variabilità delle distribuzioni a priori ai livelli superiori della gerarchia: sistemi culturali diversi hanno costruito modelli generativi diversi per i principi della distribuzione equa, e questi modelli producono giudizi di giustizia diversi in risposta agli stessi scenari. Questa variabilità non implica relatività del contenuto nel senso pernicioso: non implica che non ci sia nulla di determinato che il concetto GIUSTIZIA rappresenti. Implica che il contenuto del concetto è relativo al modello generativo del sistema che lo possiede — esattamente come il contenuto di qualsiasi altro concetto nella teleosemantica computazionale sincronica.

C'è però una differenza rilevante tra GIUSTIZIA e CAUSALITÀ o NUMERO che il framework deve affrontare esplicitamente. I concetti di CAUSALITÀ e di NUMERO hanno un referente che è in qualche misura indipendente dalle pratiche sociali: le sequenze causali e le cardinalità degli insiemi esistono indipendentemente dal fatto che comunità umane le riconoscano come tali. GIUSTIZIA, al contrario, sembra avere un referente che è costitutivamente dipendente dalle pratiche sociali e normative: una distribuzione non è giusta o ingiusta indipendentemente da qualsiasi sistema normativo, ma solo rispetto a un sistema di norme che definisce cosa conta come trattamento equo.

La risposta che la teleosemantica computazionale sincronica offre a questa difficoltà è di carattere pragmatista: la circolarità non è viziosa ma costitutiva di una classe di concetti il cui contenuto è

determinato dalla struttura delle pratiche sociali in cui sono incorporati. Nel framework predittivo, questi concetti hanno strutture generative che includono rappresentazioni di pratiche sociali come variabili causali rilevanti: il concetto GIUSTIZIA genera predizioni che dipendono non solo dalla distribuzione delle risorse ma dalla struttura delle aspettative normative della comunità in cui il giudizio viene espresso. Il grounding strutturale di GIUSTIZIA include dunque il grounding nelle pratiche sociali — nelle regolarità comportamentali e normative della comunità — che sono esse stesse parte dell'ambiente causale del sistema cognitivo.

VI.6 — Implicazioni per il problema della normatività semantica diacronica

L'analisi dell'apprendimento del modello generativo consente di affrontare un problema che la teoria sincronica del contenuto — quella sviluppata nei capitoli precedenti — non è in grado di affrontare da sola: il problema della normatività semantica diacronica, ovvero il problema di come il contenuto di una rappresentazione possa essere normativo non solo rispetto al modello generativo attuale ma rispetto alla traiettoria di sviluppo del modello attraverso il tempo.

Il problema si presenta nella forma seguente. Se il contenuto di «gatto» è determinato dalla struttura del modello generativo attuale, e se il modello di un bambino di tre anni è diverso da quello di un adulto, il bambino e l'adulto hanno rappresentazioni di gatto con contenuti diversi. Questo non è in sé problematico — la variazione del contenuto attraverso lo sviluppo è un fatto documentato dalla psicologia cognitiva dello sviluppo — ma genera una domanda: in che senso il bambino che ha una rappresentazione incorretta di gatto (che include, diciamo, anche i gatti dei cartoni animati come esemplari prototipici) *sbaglia*? Se il contenuto è determinato dal modello attuale, la rappresentazione del bambino è corretta rispetto al suo modello, anche se è diversa da quella dell'adulto.

La risposta che il framework predittivo offre a questo problema è di carattere strutturale. Il sistema di apprendimento è progettato — sia per selezione evolutiva sia per costruzione nei sistemi artificiali — per convergere verso modelli generativi che minimizzano l'energia libera su tempi lunghi e su distribuzioni di input ampie. Il modello del bambino è in uno stato transitorio che produce errori di predizione sistematici quando viene applicato a input che il modello adulto gestisce correttamente: l'incontro con nuove istanze di gatti che non corrispondono alle predizioni del modello infantile genera errori di predizione che spingono il modello a evolversi in direzione del modello adulto. La normatività diacronica non è assente dal framework ma è codificata nel processo di apprendimento stesso: la «direzione normativa» dello sviluppo del contenuto è determinata dalla traiettoria di

minimizzazione dell'energia libera su tempi lunghi, che tende verso modelli generativi sempre più adeguati rispetto alla struttura causale dell'ambiente. # VII — L'intenzionalità originaria nel framework predittivo: verso una teoria non-derivativa del contenuto

VII.1 — Il problema dell'intenzionalità originaria riformulato

Nel Capitolo I si è stabilita la distinzione tra intenzionalità originaria e intenzionalità derivata come uno dei vincoli fondamentali che qualsiasi teoria naturalistica del contenuto deve affrontare senza elusioni. La critica di Searle ai sistemi computazionali si concentra proprio su questa distinzione: i sistemi che manipolano simboli hanno al più intenzionalità derivata, una proprietà attribuita dall'esterno dagli osservatori che interpretano i loro stati secondo convenzioni consolidate. L'intenzionalità genuina — quella che caratterizza i pensieri umani e presumibilmente quelli di altri organismi biologici dotati di un sistema nervoso sufficientemente complesso — sarebbe una proprietà irriducibile della coscienza, non replicabile attraverso l'implementazione funzionale di processi computazionali per quanto raffinati.

Questa critica, applicata al framework della codifica predittiva, assume una forma specifica e merita di essere articolata prima di essere contestata. Si potrebbe sostenere che anche un sistema predittivo gerarchico, per quanto matematicamente sofisticato e neurobiologicamente plausibile nei suoi fondamenti, non faccia che implementare processi di aggiornamento statistico — manipolazione di distribuzioni di probabilità, calcolo di gradienti di energia libera, propagazione di errori di predizione — senza che nessuno di questi processi sia genuinamente *di* qualcosa nel senso brentiano del termine. Le distribuzioni a posteriori $Q^{(t)}$, le funzioni generative $g^{(t)}$ e le matrici di precisione $\Pi^{(t)}$ sarebbero strutture matematiche cui l'osservatore umano attribuisce contenuto semantico — sulla base di giudizi di analogia con il proprio caso —, ma che non lo possiedono autonomamente, originariamente, in virtù di proprietà intrinseche del sistema fisico che le implementa.

La risposta che si sviluppa nel presente capitolo non è la confutazione diretta di questa critica nella sua forma metafisica più radicale — che probabilmente eccede le possibilità di qualsiasi argomento puramente funzionale, come del resto Searle stesso ha sempre riconosciuto implicitamente adottando l'argomento del qualia come residuo irriducibile — ma la sua disarmatura strategica attraverso una distinzione che la formulazione searleana non opera con sufficiente nitidezza analitica. La questione dell'intenzionalità originaria si biforca, sotto uno sguardo più attento, in due questioni radicalmente distinte che la tradizione filosofica ha spesso confuso o deliberatamente amalgamato per ragioni

argomentative contingenti: la questione dell'ancoraggio causale del contenuto — in virtù di che cosa i processi del sistema sono *di* strutture del mondo e non di qualcos'altro o di niente — e la questione della soggettività fenomenica del contenuto — in virtù di che cosa vi è qualcosa che è *come* essere in quello stato, una prospettiva in prima persona sul contenuto che lo colora qualitativamente. La prima questione è quella a cui il framework predittivo può fornire una risposta genuinamente non-derivativa; la seconda è quella che pertiene alla coscienza fenomenica e che si affronterà nel capitolo successivo con gli strumenti concettuali appropriati. Tenere separate queste due questioni consente di vedere che la critica dell'intenzionalità derivata, nella formulazione di Searle e nelle sue varianti neodualistiche, confonde i due problemi e attribuisce alla seconda questione il peso argomentativo che solo la prima può sostenere sul piano della semantica naturalizzata.

La distinzione tra le due componenti dell'intenzionalità originaria ha implicazioni importanti per la strategia difensiva del naturalismo semantico. Il naturalismo semantico non è obbligato a risolvere il problema difficile della coscienza per stabilire una teoria del contenuto non-derivativa: può limitarsi a mostrare che gli stati del sistema hanno ancoraggio causale strutturale — che il loro contenuto è determinato dalla struttura interna del sistema in virtù di relazioni causali con il mondo — senza pretendere di spiegare perché quegli stati siano fenomenicamente caratterizzati. Questa posizione modesta ma robusta è sufficiente per rispondere alle obiezioni teleosemantiche e informative esaminate nei capitoli precedenti, e lascia aperta la questione fenomenologica per una trattazione separata che non intacca il nucleo della teoria.

VII.2 — L'ancoraggio causale come relazione strutturale interna

L'intenzionalità derivata, nel senso tecnico elaborato da Searle, è quella dei segni e dei simboli convenzionali: è derivata perché dipende dall'interpretazione di agenti esterni che attribuiscono contenuto agli stati del sistema attraverso una convenzione linguistica condivisa. La parola scritta «gatto» non è *di* gatti in virtù di alcuna proprietà intrinseca della sequenza di caratteri tipografici: è *di* gatti perché agenti linguisticamente competenti la usano per riferirsi a gatti, e questa competenza è essa stessa fondata sull'intenzionalità originaria di quegli agenti che la detengono come capacità cognitiva irriducibile. Togliere l'interpretazione degli agenti competenti, e la parola perde ogni connessione con i gatti, ridotta a pattern di inchiostro su carta o di pixel su schermo.

Nel caso di un sistema predittivo gerarchico, la relazione che connette i processi interni del sistema alle strutture del mondo è strutturalmente diversa dalla relazione convenzionale che connette la parola ai gatti, e questa differenza è di tipo, non di grado. Il modello generativo non è un insieme di regole di corrispondenza

convenzionale tra stati interni e stati del mondo: è una struttura causale complessa che si è formata attraverso l'interazione prolungata del sistema con l'ambiente e che codifica ipotesi sulla struttura causale di quell'ambiente. Le funzioni generative $\mathbf{g}^{(1)}$ non sono state definite per convenzione da nessun agente esterno — non vi è nessun progettista che abbia stabilito che $\mathbf{g}^{(1)}(\boldsymbol{\mu}^{(1)})$ debba corrispondere alla predizione di stimoli felini piuttosto che canini. Quelle funzioni hanno la struttura che hanno perché il processo di apprendimento — la minimizzazione dell'energia libera sui tempi lunghi attraverso l'esposizione all'ambiente — le ha plasmate in modo tale da riflettere la struttura causale dell'ambiente in cui il sistema è situato. La loro struttura è un rispecchiamento causale delle strutture ambientali, non un'attribuzione convenzionale dall'esterno.

Questa differenza strutturale tra dipendenza convenzionale e dipendenza causale merita di essere articolata in modo più preciso attraverso la nozione di *ancoraggio causale strutturale*. Diciamo che uno stato interno S di un sistema C ha ancoraggio causale strutturale rispetto alla proprietà F dell'ambiente se e solo se sono soddisfatte tre condizioni che si possono enunciare in modo sempre più tecnico. La prima condizione è genealogica: S fa parte di un modello generativo M che è stato plasmato causalmente dall'interazione di C con istanze di F nell'ambiente, nel senso che la struttura attuale di M non sarebbe quella che è in assenza di quella storia di interazione causale. La seconda condizione è predittiva: M genera predizioni che si accordano sistematicamente con le osservazioni prodotte da istanze di F — il modello è calibrato sull'evidenza relativa a F nel senso bayesiano di riflessione della struttura probabilistica dell'ambiente. La terza condizione è controfattuale: la struttura di accordo tra le predizioni di M e le osservazioni di F si mantiene attraverso variazioni delle condizioni di osservazione che preservano la presenza di F — il sistema non è sintonizzato su aspetti accidentali o contingenti delle istanze di F incontrate, ma sulla proprietà F stessa nella sua generalità causalmente rilevante.

Si può mostrare, con argomentazione tecnica, che il framework della codifica predittiva soddisfa queste tre condizioni quando il sistema ha appreso il modello generativo attraverso esposizione sufficiente all'ambiente. La condizione genealogica è soddisfatta per costruzione del processo di apprendimento: la regola $\boldsymbol{\theta} = -\eta \partial \mathbf{F} / \partial \boldsymbol{\theta}$ assicura che i parametri del modello siano modificati in direzione che riduce l'errore di predizione, e dunque che la struttura del modello risultante rifletta la struttura dell'ambiente che ha generato gli input osservati. Un modello che sia stato addestrato esclusivamente su gatti avrà funzioni generative che riflettono la struttura morfologica, comportamentale e contestuale dei gatti; un modello addestrato anche su cani avrà funzioni generative che riflettono anche quella struttura, con differenziazione tra i due tipi di stimoli proporzionale all'evidenza accumulata.

La condizione predittiva è soddisfatta per definizione della minimizzazione dell'energia libera: un modello che non generasse predizioni accordate con le osservazioni non minimizzerebbe F e verrebbe modificato dall'aggiornamento parametrico fino a quando non producesse predizioni più adeguate. Il punto stazionario del processo di apprendimento è, per definizione, la configurazione del modello in cui le predizioni generate sono minimamente discrepanti rispetto alle osservazioni — dunque massimamente accordate con la struttura del mondo osservato. La condizione controfattuale è soddisfatta in virtù della struttura gerarchica del modello: le rappresentazioni ai livelli superiori della gerarchia codificano strutture causali *invarianti* attraverso variazioni delle condizioni sensoriali locali. Il modello generativo a un livello sufficientemente alto non è sintonizzato sulle proprietà locali e contingenti di una singola istanza di F osservata in determinate circostanze, ma sulla struttura astratta di F che si manifesta attraverso istanze diverse, modalità diverse, contesti diversi. Questa invarianza acquisita è esattamente la robustezza controfattuale richiesta dalla terza condizione.

VII.3 — La critica del regresso dell'interpretazione

Un argomento complementare che rafforza la posizione dell'ancoraggio causale non-derivativo riguarda il problema del regresso cui va incontro qualsiasi teoria che derivi l'intenzionalità dall'interpretazione. Se il contenuto di S dipende dall'interpretazione di un agente A , allora il contenuto di S non è determinato dalle proprietà intrinseche di S ma dalle proprietà intrinseche di A — che include, come componente essenziale, gli stati intenzionali di A che gli consentono di interpretare S nel modo appropriato. Ma se l'intenzionalità di A è essa stessa derivata — se dipende a sua volta dall'interpretazione di un ulteriore agente A' —, si genera un regresso infinito che non determina mai il contenuto. Il regresso si ferma solo se esiste almeno un livello di intenzionalità che sia genuinamente originaria — non dipendente dall'interpretazione di nessun agente esterno. La proposta della teleosemantica computazionale sincronica è che l'ancoraggio causale strutturale fornisce precisamente quel livello: il contenuto degli stati di un sistema con modello generativo adeguatamente plasmato dall'ambiente è determinato dalle proprietà intrinseche del sistema e dalla sua storia causale, senza dipendere dall'interpretazione di nessun agente ulteriore.

Questo argomento anti-regressivo non pretende di mostrare che il sistema predittivo abbia la stessa qualità di intenzionalità di un essere umano adulto: si limita a mostrare che il sistema ha intenzionalità non-derivativa nel senso rilevante per la determinazione del contenuto. Le qualità fenomeniche dell'intenzionalità umana — la prospettiva soggettiva, la colorazione affettiva degli stati, il carattere di «minezza» dell'esperienza — possono essere proprietà aggiuntive che il sistema predittivo possiede solo in gradi variabili a seconda del tipo

di implementazione biologica e della complessità del modello di sé. Ma queste qualità non sono necessarie per la non-derivatività del contenuto nel senso in cui la teoria semantica la richiede: esse riguardano la fenomenologia dell'intenzionalità, non la sua struttura semantica.

VII.4 — Intenzionalità originaria debole e forte: una tassonomia più precisa

L'ancoraggio causale strutturale soddisfa una nozione di intenzionalità originaria che si propone di chiamare *debole*: il contenuto degli stati del sistema è determinato dalla struttura interna del sistema stesso, senza dipendere dall'attribuzione di un agente interpretante esterno. Questa nozione è sufficiente per rispondere all'obiezione searleana nella sua formulazione standard — che i sistemi computazionali abbiano solo intenzionalità derivata nel senso in cui la hanno le parole scritte — ma non è sufficiente per rispondere alla versione più radicale dell'obiezione, che richiede non solo che il contenuto sia internamente determinato ma che vi sia qualcosa che è *come* avere quello stato per il sistema stesso.

L'intenzionalità originaria *forte* richiede questa componente fenomenica aggiuntiva: non basta che gli stati del sistema abbiano contenuto determinato internamente, occorre che siano stati sentiti dal sistema come stati di un certo tipo — che vi sia una qualità intrinseca dell'esperienza di quell' stato, un certo sapore fenomenico. Questa forma più esigente di intenzionalità è ciò che caratterizza intuitivamente la differenza tra la percezione visiva di un gatto da parte di un essere umano e l'elaborazione di un pattern visivo coerente con un gatto da parte di un sistema di visione artificiale altamente sofisticato. Entrambi i sistemi possono avere stati con contenuto determinato internamente nel senso debole; solo il primo ha — presumibilmente, ammettendo una buona dose di incertezza epistemica sul problema difficile — stati con contenuto fenomenicamente caratterizzato nel senso forte.

La questione dell'intenzionalità forte è inseparabile dalla questione della coscienza fenomenica, e si affronterà nel capitolo successivo con strumenti concettuali appropriati. Quello che importa stabilire in questa sede è che la distinzione tra intenzionalità forte e debole non invalida la proposta sviluppata nei capitoli precedenti: anche se il framework predittivo non fornisce da solo una teoria completa dell'intenzionalità forte — che richiederebbe una soluzione al problema difficile della coscienza —, fornisce una teoria robusta dell'intenzionalità debole. Questa teoria risponde in modo non-elusivo alle obiezioni delle teorie concorrenti esaminate nei capitoli II, III e V: risponde alla teleosemantica millikaniana con una struttura normativa non-diacronica, alla semantica informazionale dretsiana con una dissoluzione del problema della disgiuntività, e al fodorismo con un account della composizionalità che non sacrifica il contestualismo. Le questioni che rimangono

aperte riguardano la coscienza come tale, non il contenuto, e richiedono strumenti distinti che non mettono in discussione i risultati ottenuti.

C'è una terza nozione che merita di essere introdotta nella tassonomia: quella di intenzionalità originaria *intermedia*, che caratterizza i sistemi che hanno qualche forma di esperienza soggettiva ma non la ricchezza qualitativa dell'esperienza umana. Molti organismi biologici — mammiferi, uccelli, forse cefalopodi — sembrano avere esperienze soggettive sufficientemente ricche da attribuire loro una forma di intenzionalità più che meramente debole, pur senza che sia chiarito se possiedano tutta la struttura fenomenologica dell'esperienza umana. Il framework predittivo può caratterizzare questa gradazione: sistemi con modelli di sé più semplici avranno forme più elementari di intenzionalità intermedia, sistemi con modelli di sé più complessi avranno forme più ricche. La gradazione dell'intenzionalità lungo la scala evolutiva è una previsione naturale del framework che la teleosemantica millikaniana tende a binarizzare troppo bruscamente.

VII.5 — La misrappresentazione sistematica e il problema delle illusioni cognitive

Una delle previsioni più interessanti della teoria del contenuto basata sulla codifica predittiva è la spiegazione delle illusioni cognitive come forme di misrappresentazione sistematica che emergono non da malfunzionamenti del sistema ma dalla struttura ottimale del modello generativo in condizioni di stimolazione atipiche. Le illusioni percettive — l'illusione di Müller-Lyer, il cubo di Necker, le illusioni di movimento di Ebbinghaus e di Ponzo, le illusioni di dimensione del sole all'orizzonte — sono stati in cui il sistema produce rappresentazioni che non corrispondono alle proprietà fisiche degli stimoli pur essendo ottimali rispetto alla struttura del modello generativo. Non si tratta di errori del sistema — il sistema funziona correttamente minimizzando l'energia libera data la struttura del proprio modello — ma di mismatch tra la struttura del modello generativo, forgiata dall'esposizione a ambienti ecologicamente normali, e la struttura degli stimoli artificiali che producono l'illusione in condizioni di laboratorio.

L'illusione di Müller-Lyer, in cui due segmenti di lunghezza fisica identica appaiono di lunghezza diversa a seconda della direzione delle frecce aggiunte alle loro estremità, può essere spiegata nei termini del framework predittivo in modo che illustra con chiarezza la struttura della misrappresentazione sistematica. Il modello generativo del sistema visivo umano ha appreso, attraverso l'esposizione prolungata ad ambienti tridimensionali con strutture angolari — spigoli, angoli, prospettive architettoniche —, che frecce convergenti sono associate statisticamente ad angoli sporgenti che si allontanano dall'osservatore, mentre frecce divergenti sono associate ad angoli rientranti che si avvicinano. In un ambiente tridimensionale, questo apprendimento è adattivo: la

lunghezza percepita di un segmento deve essere corretta per la prospettiva, e la direzione delle frecce fornisce informazione sulla prospettiva che consente la costanza delle dimensioni — la capacità di percepire oggetti della stessa dimensione come ugualmente grandi indipendentemente dalla distanza.

Quando lo stesso tipo di stimolo viene presentato su una superficie bidimensionale — dove la correzione prospettica è inappropriata perché non vi è struttura tridimensionale reale — il modello generativo applica ugualmente la correzione prospettica, producendo la distorsione percettiva caratteristica dell'illusione. Il sistema non sbaglia perché funziona male: sbaglia perché funziona bene in un contesto per cui non è stato «progettato» — un contesto artificiale che sfrutta sistematicamente le strutture del modello per produrre inferenze erranee. L'illusione rivela la struttura del modello generativo, non un difetto del sistema inferenziale.

Formalmente, la rappresentazione della lunghezza del segmento al livello 2 della gerarchia visiva è determinata non solo dall'input sensoriale al livello 0 ma anche dalle predizioni top-down generate dalla rappresentazione della prospettiva al livello 3. Sia $P^{(3)}$ la rappresentazione della prospettiva e $L^{(2)}$ la rappresentazione della lunghezza del segmento. Il modello generativo specifica una funzione $g^{(3)}$ che dipende dal valore di $P^{(3)}$:

$g^{(3)}(P^{(3)}) = \text{convergente}$, $L^{(2)} = \text{predizione di stimolo visivo come per segmento più lungo}$
 $g^{(3)}(P^{(3)}) = \text{divergente}$, $L^{(2)} = \text{predizione di stimolo visivo come per segmento più corto}$

La rappresentazione effettiva di $L^{(2)}$ è determinata dal bilancio tra il segnale bottom-up — che indica la lunghezza fisica — e il segnale top-down — che indica la lunghezza predetta data la prospettiva percepita. La ponderazione tra questi due segnali è regolata dalle matrici di precisione: quando la precisione delle predizioni prospettiche è alta rispetto alla precisione del segnale sensoriale locale, la rappresentazione di $L^{(2)}$ è dominata dalla predizione prospettica e si discosta dalla lunghezza fisica effettiva.

Questa formalizzazione spiega una caratteristica fenomenica notevole delle illusioni di tipo Müller-Lyer: esse sono robuste anche quando l'osservatore sa intellettualmente che i segmenti sono uguali. La conoscenza intellettuale della situazione modifica le credenze ai livelli gerarchici superiori — al livello del ragionamento esplicito, il soggetto rappresenta correttamente che i segmenti hanno la stessa lunghezza —, ma non modifica la precisione dei segnali di prospettiva ai livelli intermedi della gerarchia, che continuano a dominare la rappresentazione percettiva immediata della lunghezza. La dissociazione tra conoscenza intellettuale e percetto immediato è una predizione strutturale del framework, non un esito contingente: i livelli superiori della gerarchia possono contenere credenze che conflittano con le predizioni dei livelli intermedi, e la

rappresentazione fenomenica immediata è determinata dai livelli intermedi più che dai livelli superiori, per ragioni di architettura del sistema inferenziale.

Questa spiegazione delle illusioni ha implicazioni importanti per la teoria della normatività del contenuto. Le illusioni non sono stati in cui il sistema va *contro* la propria funzione normativa: sono stati in cui la funzione normativa è implementata correttamente ma produce risultati che divergono dalla realtà fisica in contesti atipici. La normatività del sistema predittivo è calibrata sull'ambiente ecologicamente normale, non su tutti i contesti possibili; le illusioni emergono precisamente quando il contesto si discosta dall'ambiente di calibrazione. Questa struttura è esattamente quella che ci si aspetta da un sistema ottimale bayesiano: un sistema ottimale non è infallibile, è ottimale rispetto a una distribuzione a priori sull'ambiente, e produce sistematicamente errori quando l'ambiente si discosta da quella distribuzione.

VII.6 — Intenzione, azione e il problema dell'agentività

Il framework predittivo, nella sua versione estesa all'azione sviluppata da Friston e collaboratori a partire dal 2010 — il framework dell'inferenza attiva —, include non solo la percezione ma anche l'azione come processo di minimizzazione dell'energia libera variazionale. Nell'inferenza attiva, l'agente non si limita ad aggiornare le proprie distribuzioni a posteriori sullo stato del mondo per accordarle con le osservazioni: può anche agire sull'ambiente per modificare le osservazioni stesse in modo che si accordino con le proprie predizioni. Percezione e azione sono dunque due modalità complementari e concorrenti di riduzione dell'errore di predizione: la percezione riduce l'errore aggiornando le credenze sul mondo, l'azione riduce l'errore modificando il mondo affinché corrisponda alle credenze. Questa simmetria è concettualmente profonda: suggerisce che la distinzione tra agente e ambiente non sia assoluta ma dipenda dal modo in cui il sistema sceglie di gestire l'errore di predizione.

Formalmente, il sistema di inferenza attiva agisce su variabili di azione $\mathbf{a}$ che influenzano le osservazioni sensoriali attraverso la funzione di accoppiamento $\mathbf{h}(\mathbf{s}, \mathbf{a})$: le osservazioni dipendono non solo dallo stato del mondo $\mathbf{s}$ ma anche dalle azioni dell'agente, secondo il modello:

$$\mathbf{o} = \mathbf{h}(\mathbf{s}, \mathbf{a}) + \boldsymbol{\varepsilon}_{\text{obs}}$$

dove $\boldsymbol{\varepsilon}_{\text{obs}}$ è il rumore di osservazione. In questo contesto esteso, l'energia libera variazionale diventa funzione sia dei parametri variazionali $\boldsymbol{\mu}$ sia delle variabili di azione $\mathbf{a}$:

$$\mathbf{F}[\mathbf{Q}, \mathbf{o}(\mathbf{a})] = D_{\text{KL}}[\mathbf{Q}(\mathbf{s}; \boldsymbol{\mu}) \parallel \mathbf{P}(\mathbf{s})] - E_{\mathbf{Q}}[\log \mathbf{P}(\mathbf{o}(\mathbf{a}) \mid \mathbf{s})]$$

La minimizzazione di $\mathbf{F}$ rispetto ad $\mathbf{a}$ — mantenendo fissi i parametri variazionali $\boldsymbol{\mu}$ che specificano le predizioni sugli stati del mondo — produce la regola di azione:

$$\dot{\mathbf{a}} = -\kappa \partial \mathbf{F} / \partial \mathbf{a} = \kappa (\partial \mathbf{o} / \partial \mathbf{a})^T \Pi^{(0)} \epsilon^{(0)}$$

dove κ è la costante di velocità dell'azione, il fattore $(\partial \mathbf{o} / \partial \mathbf{a})^T$ è il jacobiano trasposto della funzione di accoppiamento azione-osservazione, e $\epsilon^{(0)} = \mathbf{o} - \hat{\mathbf{g}}^{(0)}(\boldsymbol{\mu}^{(0)})$ è l'errore di predizione sensoriale al livello basale. Questa equazione descrive l'azione come un processo di riduzione dell'errore di predizione sensoriale attraverso la modifica delle osservazioni: l'agente si muove in modo da portare le proprie osservazioni in accordo con le proprie predizioni top-down. L'azione non è guidata da «desire» e «belief» separati nel senso della psicologia del senso comune, ma dalla struttura dell'errore di predizione attuale rispetto al modello predittivo in atto.

La rilevanza di questa struttura per la questione dell'intenzionalità originaria è diretta e merita di essere sottolineata. L'intenzione, nel framework dell'inferenza attiva, non è una causa che precede e produce l'azione dall'esterno del sistema computazionale — come un fantasma nella macchina che attiva il meccanismo corporeo — ma è una rappresentazione di stati futuri desiderati: stati che il sistema predice come configurazioni di minima energia libera verso cui orientare la propria attività. Le intenzioni sono predizioni su stati futuri propri; l'azione è il meccanismo che realizza quelle predizioni riducendo la discrepanza tra gli stati predetti come obiettivi e gli stati attualmente osservati. Questa struttura dissolve la dicotomia tra causa mentale e causa fisica dell'azione — uno dei nodi centrali del problema della causazione mentale esaminato nel Capitolo I — non separando i due domini ma mostrando che essi sono aspetti complementari del medesimo processo di minimizzazione dell'energia libera.

La questione dell'agentività — che cosa distingue un agente genuino da un sistema che meramente reagisce agli stimoli — riceve nel framework dell'inferenza attiva una risposta precisa. Un agente genuino è un sistema che non si limita a ricevere passivamente l'input dall'ambiente e ad aggiornare le proprie credenze, ma che genera predizioni su stati futuri propri — stati *goal* — e che seleziona azioni sulla base della discrepanza tra gli stati attuali e gli stati goal predetti. La caratteristica distintiva dell'agentività non è la presenza di un homunculus interno che decide liberamente, ma la struttura gerarchica del modello predittivo: i livelli superiori della gerarchia generano predizioni su stati futuri dell'agente — intenzioni — che i livelli intermedi e bassi cercano di realizzare attraverso la selezione delle azioni appropriate. L'agentività emerge dalla profondità temporale del modello predittivo: un sistema che predicesse solo il futuro immediato sarebbe meramente reattivo; un sistema che predicesse traiettorie temporalmente estese verso stati futuri desiderati sarebbe genuinamente agente.

VII.7 — Intenzioni collettive e intenzionalità distribuita

Un'estensione del framework che merita attenzione riguarda la possibilità di dare conto dell'intenzionalità collettiva — delle intenzioni che i gruppi di agenti hanno *come gruppi*, non semplicemente come aggregati di intenzioni individuali. La questione dell'intenzionalità collettiva ha ricevuto trattazione dettagliata da Searle, Bratman e Tuomela nel contesto della filosofia dell'azione sociale, con enfasi sulla non-riducibilità delle intenzioni collettive alle intenzioni individuali in modo semplice.

Nel framework predittivo esteso, l'intenzionalità collettiva può essere concepita come l'allineamento dei modelli predittivi di agenti multipli in modo tale che le azioni di ciascuno riducono gli errori di predizione di tutti. Quando un gruppo di agenti ha modelli generativi parzialmente condivisi — costruiti attraverso un'esperienza comune, comunicazione linguistica e coordinazione comportamentale —, le loro predizioni tendono a convergere su stati futuri simili e le loro azioni tendono a coordinarsi per realizzare quegli stati. L'intenzione condivisa non è una misteriosa entità sovra-individuale: è la struttura di accoppiamento tra i modelli generativi dei singoli agenti che rende le loro predizioni e azioni mutuamente coerenti.

La comunicazione linguistica svolge un ruolo cruciale in questo processo di allineamento dei modelli. Il linguaggio non è solo un mezzo di trasmissione di informazioni già codificate nei modelli individuali: è un meccanismo attivo di modifica dei modelli generativi degli interlocutori — un meccanismo di accoppiamento tra modelli predittivi che riduce la divergenza KL tra le distribuzioni dei diversi agenti. Quando un parlante produce un enunciato, modifica le predizioni del suo interlocutore su stati futuri del mondo e degli agenti, allineando i modelli generativi dei due soggetti in modo che le loro azioni future siano più facilmente coordinabili. La pragmatica del linguaggio — lo studio di come gli enunciati modificano le credenze e le intenzioni degli interlocutori — è, nel framework predittivo, la teoria di come la comunicazione modifica le distribuzioni a posteriori dei partecipanti all'interazione comunicativa.

VIII — Contenuto fenomenico e modello di sé: la coscienza come struttura predittiva di ordine superiore

VIII.1 — Il problema difficile della coscienza e il suo rapporto con il contenuto semantico

La distinzione tra il problema facile e il problema difficile della coscienza, proposta da David Chalmers nel 1995 in un articolo diventato immediatamente uno dei riferimenti centrali della filosofia della mente contemporanea, ha riconfigurato il dibattito sulla mente in modo che continua a essere profondamente influente. I problemi facili riguardano l'esplicazione delle capacità cognitive — la discriminazione percettiva, l'integrazione dell'attenzione, la reportabilità verbale degli stati mentali, la regolazione e il controllo del comportamento, il ciclo sonno-veglia — e ammettono, in linea di principio, una soluzione funzionale: è possibile specificare meccanismi computazionali che realizzino quelle capacità in modo soddisfacente. Il problema difficile riguarda invece l'esistenza dell'esperienza soggettiva stessa — il fatto che l'elaborazione dell'informazione sia accompagnata da qualcosa che è *come* effettuarla, da una prospettiva in prima persona, da qualia fenomenici che colorano qualitativamente le esperienze. Questo fatto non sembra spiegabile, nemmeno in linea di principio, attraverso la sola specificazione funzionale: anche se si descrivesse completamente il meccanismo computazionale della percezione del colore rosso, rimarrebbe aperta la domanda perché quell'elaborazione sia accompagnata dall'esperienza soggettiva del rosso — dalla *redness* nella sua qualità immediata e irreducibile.

Il rapporto tra il problema della coscienza fenomenica e il problema del contenuto semantico è più stretto di quanto la distinzione di Chalmers lasci apparire, e questa stretta interdipendenza ha generato confusione in entrambe le direzioni. Da un lato, il problema difficile della coscienza può essere visto come un aspetto specifico del problema dell'intenzionalità originaria forte: l'intenzionalità fenomenica — il fatto che gli stati coscienti siano *di* qualcosa e che questo essere-*di* sia fenomenicamente caratterizzato — è un aspetto del problema difficile che riguarda tanto la semantica quanto la fenomenologia. Dall'altro lato, alcune soluzioni proposte al problema del contenuto hanno implicazioni dirette per il problema della coscienza: se il contenuto è puramente funzionale, si può essere tentati di concludere che anche la coscienza è puramente funzionale, il che conflitterebbe con le intuizioni robuste che alimentano il problema difficile nella sua formulazione più radicale.

La posizione sviluppata nel presente capitolo è che il framework della codifica predittiva offre una caratterizzazione computazionalmente precisa di alcuni aspetti della coscienza fenomenica — in particolare, della struttura temporale del flusso di

coscienza, della coscienza di sé, dell'unità dell'esperienza e delle alterazioni correlate con le sostanze psicotrope — senza pretendere di risolvere il problema difficile nel senso di spiegare perché vi sia qualcosa che è come essere un sistema predittivo. Questa modestia epistemica non è una resa intellettuale o una defezione di fronte al problema più duro: è la conseguenza di una distinzione precisa tra ciò che il framework può spiegare nella sua forma computazionale e ciò che rimane, almeno provvisoriamente, fuori dalla portata di qualsiasi spiegazione puramente funzionale. Comprendere con precisione qual è la struttura computazionale dell'esperienza cosciente è un passo necessario — e forse il più importante di cui disponiamo attualmente — verso qualsiasi risposta al problema difficile, anche se non è di per sé quella risposta.

VIII.2 — La coscienza come inferenza sugli stati propri: il modello di sé

La proposta centrale del presente capitolo è che la coscienza fenomenica, nella misura in cui è accessibile a una caratterizzazione computazionale senza presupporre la soluzione del problema difficile, corrisponde a un *modello generativo di sé* che la gerarchia predittiva costruisce e aggiorna continuamente nel corso dell'esperienza. Il modello di sé non è un livello separato del sistema cognitivo, un'istanza di meta-cognizione che si sovrappone alla cognizione ordinaria: è il modo in cui la gerarchia predittiva si rappresenta essa stessa come agente situato, che si trova in un ambiente e che interagisce con esso in modo orientato al futuro. Il modello di sé è, in termini formali, la componente del modello generativo complessivo che rappresenta l'agente stesso come variabile causale nel processo di generazione delle osservazioni — non solo come sistema ricevente che elabora input ma come sorgente attiva di azioni che influenzano le osservazioni future.

Definiamo formalmente il modello di sé come la distribuzione a posteriori sugli stati dell'agente stesso — stati propriocettivi, stati fisiologici interni, stati di azione intenzionale, stati cognitivi di ordine superiore — che la gerarchia predittiva mantiene in parallelo con la distribuzione sugli stati dell'ambiente esterno. Sia $\mathbf{s_self}$ il vettore degli stati propri dell'agente e $\mathbf{s_world}$ il vettore degli stati del mondo esterno. Il modello generativo completo specifica una distribuzione congiunta:

$$P(\mathbf{o}, \mathbf{s_self}, \mathbf{s_world}) = P(\mathbf{o} \mid \mathbf{s_self}, \mathbf{s_world}) \cdot P(\mathbf{s_self} \mid \mathbf{s_world}) \cdot P(\mathbf{s_world})$$

dove la dipendenza di $\mathbf{o}$ sia da $\mathbf{s_self}$ sia da $\mathbf{s_world}$ riflette il fatto che le osservazioni sensoriali dipendono sia dalle strutture del mondo esterno sia dagli stati dell'agente che fa le osservazioni — le osservazioni visive dipendono non solo dagli oggetti presenti nello spazio ma anche dalla posizione degli occhi,

dall'orientamento della testa, dallo stato delle pupille e dalle saccadi oculari.

La distribuzione a posteriori $Q(\mathbf{s}_{\text{self}}; \boldsymbol{\mu}_{\text{self}})$ — la stima a posteriori degli stati propri dell'agente — costituisce il modello di sé in senso stretto. Il contenuto del modello di sé è determinato dalla medesima struttura triadica identificata per le rappresentazioni dell'ambiente: la distribuzione a posteriori sugli stati propri, la funzione generativa che mappa gli stati propri nelle predizioni sensoriali propriocettive, e la matrice di precisione che determina il peso attentivo attribuito ai segnali propriocettivi. Questa struttura omologa tra il modello del mondo e il modello di sé non è accidentale: riflette il fatto che il sistema cognitivo si rappresenta sé stesso con gli stessi strumenti con cui si rappresenta l'ambiente, secondo un principio di simmetria che ha implicazioni teoriche profonde per la comprensione dell'autoconsapevolezza.

VIII.3 — La struttura temporale della coscienza: ritenzione, presentazione e protenzione

Uno degli aspetti più teoricamente ricchi del modello di sé nel framework predittivo è la sua struttura temporale, che corrisponde con precisione sorprendente alla struttura fenomenologica del flusso di coscienza che Edmund Husserl aveva analizzato nelle sue *Lezioni sulla fenomenologia della coscienza interna del tempo* tenute tra il 1905 e il 1917 e pubblicate postumanamente. L'analisi husserliana è tra le più sottili e tecnicamente impegnative dell'intera tradizione fenomenologica, e vale la pena ricostruirla prima di mostrarne la corrispondenza computazionale.

Husserl osservava che la percezione del presente non è puntuale — non è un accesso istantaneo e senza durata a un «adesso» isolato —, ma è temporalmente estesa in un modo che non richiede la riflessione per essere esperita. La coscienza temporale primaria include tre momenti che non sono successivi ma strutturalmente contemporanei nell'atto percettivo: la ritenzione (*Retention*), che è la coscienza del passato immediatamente trascorso ancora vivo nella percezione presente — il «ancora udito» di una nota musicale appena cessata che continua a determinare il significato della nota presente —; la presentazione vivente (*lebendige Gegenwart*), che è il momento di attenzione immediata a ciò che accade ora —; e la protenzione (*Protention*), che è l'anticipazione del futuro immediatamente imminente già presente nella struttura dell'atto percettivo, il «già atteso» della nota che seguirà. Questa struttura tripartita non è un artificio della riflessione consapevole: è costitutiva della percezione stessa, del modo in cui ogni momento percettivo porta in sé le tracce del passato e l'anticipazione del futuro.

Il framework predittivo formalizza questa struttura fenomenologica in modo computazionalmente preciso. La distribuzione a posteriori che il sistema mantiene non è distribuita

sugli stati istantanei dell'ambiente in un singolo momento ma sulle *traiettorie* degli stati attraverso il tempo — sulle sequenze di stati passati, presenti e futuri che il modello generativo considera probabili data la struttura della dinamica ambientale. Il modello generativo temporale specifica una distribuzione sulle traiettorie $\mathbf{s}(t)$ per t appartenente all'intervallo $[t_{\text{prev}}, t_{\text{pred}}]$:

$$P(\mathbf{o}(t_0:t_1), \mathbf{s}(t_0:t_1)) = P(\mathbf{o}(t_0:t_1) \mid \mathbf{s}(t_0:t_1)) \cdot P(\mathbf{s}(t_0:t_1))$$

dove $P(\mathbf{s}(t_0:t_1))$ è la distribuzione a priori sulle traiettorie degli stati, tipicamente specificata attraverso un modello dinamico della forma:

$$d\mathbf{s}/dt = \mathbf{f}(\mathbf{s}, \mathbf{v}) + \boldsymbol{\omega}$$

con $\mathbf{f}$ la funzione di deriva (*drift*), $\mathbf{v}$ le cause esogene che influenzano la dinamica, e $\boldsymbol{\omega}$ il rumore di processo che introduce variabilità. In questo modello, il processo di inferenza a posteriori determina non solo la stima degli stati al tempo t corrente ma la distribuzione sull'intera traiettoria — stima degli stati passati e previsione degli stati futuri — data l'intera sequenza di osservazioni fino a t .

La distribuzione a posteriori sulle traiettorie passate — la stima del percorso seguito dagli stati fino al momento presente — corrisponde alla ritenzione husserliana: è la formalizzazione della traccia degli eventi passati ancora attiva nella determinazione della stima presente. Il punto di massima probabilità a posteriori nell'istante t corrente corrisponde alla presentazione vivente. La distribuzione predittiva sulle traiettorie future — la stima bayesiana di come gli stati evolveranno a partire dallo stato corrente — corrisponde alla protenzione. La corrispondenza tra la struttura fenomenologica e la struttura computazionale non è puramente analogica: nell'inferenza sulle traiettorie, la distribuzione a posteriori al tempo t dipende sia dagli input sensoriali ricevuti fino a t — che determinano la forma della distribuzione sulle traiettorie passate tramite l'aggiornamento bayesiano — sia dalla struttura del modello dinamico — che determina la distribuzione predittiva sulle traiettorie future attraverso la funzione di deriva $\mathbf{f}$. Il passato e il futuro non si fondono nel presente per un atto di sintesi successivo: sono già presenti, strutturalmente, in ogni stima a posteriori che il sistema calcola.

VIII.4 — L'unità dell'esperienza come coerenza del modello generativo

Il problema dell'unità della coscienza — che cosa garantisce che l'esperienza cosciente in un dato momento sia una, unitaria, coerente, piuttosto che una molteplicità di frammenti slegati — è stato formulato da Kant come il problema dell'appercezione trascendentale: l'unità del soggetto pensante come condizione di possibilità dell'esperienza unificata. In una formulazione più empirica, lo stesso problema si traduce nel cosiddetto *binding*

problem neuroscientifico: come le informazioni elaborate in parallelo da aree cerebrali diverse vengono integrate in un'esperienza unitaria e coerente di oggetti e scene percettive.

La risposta standard, che ha ispirato molte linee di ricerca sperimentale a partire dagli anni novanta, ipotizza che il *binding* sia realizzato attraverso la sincronizzazione oscillatoria nella banda gamma: le aree che elaborano aspetti diversi dello stesso oggetto — colore, forma, posizione, texture — sincronizzano le proprie oscillazioni, e questa coerenza temporale costituisce il «legame» che integra le rappresentazioni parziali in un'esperienza unitaria. L'evidenza empirica per questa ipotesi è stata a lungo controversa e i risultati sperimentali non hanno avuto la robustezza e la replicabilità che ci si attendeva, lasciando aperta la questione sul meccanismo del binding.

Il framework predittivo offre una risposta alternativa che non dipende dalla sincronizzazione oscillatoria ma dalla struttura gerarchica del modello generativo. L'unità dell'esperienza è garantita dall'unità del modello generativo che produce le predizioni per tutti i livelli della gerarchia: poiché tutte le rappresentazioni distribuite nelle diverse aree corticali sono generate come componenti di un unico processo di ottimizzazione variazionale — la minimizzazione dell'energia libera rispetto al modello generativo condiviso —, esse sono intrinsecamente coerenti fin dall'inizio. Non si tratta di integrare rappresentazioni inizialmente separate attraverso un meccanismo di legame successivo: non vi è separazione genuina, perché le rappresentazioni ai diversi livelli gerarchici sono generate dallo stesso processo di ottimizzazione e sono mutualmente vincolate attraverso la struttura delle funzioni generative che li connettono.

Formalmente, la coerenza dell'esperienza in un dato momento t è la coerenza della distribuzione a posteriori congiunta

$$Q(\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \dots, \mathbf{s}^{(L)}; \boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \dots, \boldsymbol{\mu}^{(L)})$$

rispetto all'approssimazione variazionale fattorizzata $\prod Q^{(l)}(\mathbf{s}^{(l)}; \boldsymbol{\mu}^{(l)})$. La coerenza è massimale quando l'approssimazione fattorizzata è vicina alla vera distribuzione a posteriori congiunta — quando le dipendenze tra livelli gerarchici sono catturate adeguatamente dalle correlazioni indotte dalle funzioni generative $\mathbf{g}^{(l)}$. La struttura gerarchica delle funzioni generative è il meccanismo che garantisce questa coerenza: ogni livello è vincolato dai livelli adiacenti attraverso le equazioni di aggiornamento

$$\boldsymbol{\mu}^{(l)} = (\partial \mathbf{g}^{(l)} / \partial \boldsymbol{\mu}^{(l)})^T \boldsymbol{\xi}^{(l)} - \boldsymbol{\xi}^{(l+1)}$$

e il punto stazionario dell'intero sistema è una configurazione in cui tutte le rappresentazioni ai diversi livelli sono mutuamente coerenti — una configurazione in cui nessun livello produce predizioni sistematicamente incompatibili con le osservazioni degli altri livelli.

Il rapporto tra questa coerenza computazionale e l'unità fenomenica dell'esperienza è, naturalmente, un rapporto tra correlato funzionale e proprietà fenomenologica, non un'identità riduttiva. La coerenza computazionale del modello generativo è la struttura funzionale che corrisponde — nel sistema implementato biologicamente nel cervello — all'unità fenomenica dell'esperienza cosciente. Che questa corrispondenza sia di identità, sopravvenienza, o realizzazione multipla dipende dalla metafisica della mente che si adotta, questione che il framework predittivo può lasciare aperta senza perdere il proprio valore esplicativo.

VIII.5 — Il senso di sé minimale e il sé narrativo

La distinzione tra il sé *narrativo* e il sé *minimale*, elaborata da Dan Zahavi nel contesto di una fenomenologia che dialoga con le neuroscienze, ha ricevuto supporto empirico da studi su pazienti con amnesie severe, stati vegetativi, disturbi del senso di sé come la depersonalizzazione e la schizofrenia. Il sé narrativo è la storia autobiografica coerente che un soggetto costruisce di sé stesso come agente con una storia personale, valori, relazioni, progetti futuri — la struttura diacronica dell'identità personale nel senso di Locke e Parfit. Il sé minimale è il senso immediato, preriflessivo, di essere il soggetto delle proprie esperienze presenti — il fatto che ogni esperienza sia già sempre mia, che vi sia un punto di vista in prima persona dal quale è percepita, indipendentemente da qualsiasi narrazione autobiografica.

Nel framework predittivo, questa distinzione corrisponde a due livelli gerarchici della gerarchia predittiva che hanno caratteristiche computazionali ben distinte. Il sé minimale corrisponde alla componente del modello di sé che rappresenta l'agente come sorgente prospettica delle proprie osservazioni sensoriali — la componente che include la posizione corporea, l'orientamento sensoriale, gli stati propriocettivi immediati. Il sé narrativo corrisponde alle rappresentazioni ai livelli superiori della gerarchia che integrano informazioni temporalmente estese attraverso l'autobiografia, costruiscono modelli degli stati mentali propri e altrui (la base computazionale della teoria della mente), e generano predizioni su traiettorie di vita e valori di lungo periodo.

Il sé minimale è incarnato e situato in modo immediato: dipende dalla posizione del corpo nell'ambiente, è aggiornato continuamente dagli input propriocettivi e vestibolari, e varia con ogni movimento del corpo. Non è un oggetto di riflessione ma la struttura della prima persona — il fatto che le osservazioni sensoriali siano *prospettiche*, cioè dipendano dalla posizione di un corpo che si trova in un luogo specifico dello spazio. La formalizzazione di questa struttura prospettica è la dipendenza delle osservazioni da s_{self} nella distribuzione generativa: $P(o | s_{self}, s_{world})$ codifica il fatto che le stesse strutture del mondo producono osservazioni diverse a seconda della posizione e dell'orientamento del corpo.

Il sé narrativo è costitutivamente temporale: dipende dalla ritenzione in memoria dell'intera storia biografica, è aggiornato dagli eventi significativi della vita, e si proietta in futuro attraverso piani e progetti. La sua rappresentazione ai livelli superiori della gerarchia predittiva è di carattere astratto e simbolico: include concetti come «io sono un medico», «io preferisco la musica classica al jazz», «io ho paura dell'altezza» — rappresentazioni di proprietà stabili dell'agente che generano predizioni sulle proprie risposte future a situazioni rilevanti. L'interazione tra sé minimale e sé narrativo è la struttura computazionale della revisione dell'autocomprensione: quando le risposte effettive del sé minimale — le reazioni immediate, gli stati affettivi spontanei — divergono dalle predizioni del sé narrativo, si generano errori di predizione che spingono all'aggiornamento di uno dei due livelli.

VIII.6 — Coscienza e precisione: il ruolo dei sistemi neuromodulatori

Il collegamento tra il parametro di precisione nel framework predittivo e i sistemi neuromodulatori cerebrali — dopamina, acetilcolina, noradrenalina, serotonina — è uno dei punti di contatto più robusti e fecondi tra il modello computazionale e l'evidenza neurobiologica disponibile. Friston e collaboratori hanno proposto, in una serie di articoli dal 2009 in poi, che i diversi sistemi neuromodulatori codifichino diverse componenti della struttura di precisione del sistema predittivo. La dopamina codificherebbe la precisione degli errori di predizione nelle proiezioni striatali e mesocorticali, rilevante per l'apprendimento per rinforzo, la selezione dell'azione e la modulazione del comportamento diretto a obiettivi. L'acetilcolina codificherebbe la precisione degli input sensoriali rispetto alle predizioni top-down nelle aree sensoriali primarie, rilevante per l'attenzione sensoriale selettiva e l'affinamento percettivo. La noradrenalina codificherebbe la precisione globale del sistema o l'incertezza sullo stato ambientale, rilevante per la regolazione del bilancio tra esplorazione e sfruttamento e per la risposta alle violazioni delle aspettative. La serotonina modulerebbe la precisione delle rappresentazioni degli stati affettivi e del sé, con implicazioni per la regolazione dell'umore e per la stabilità del modello di sé.

La rilevanza di questa corrispondenza per il problema della coscienza è diretta. Se la precisione determina il peso relativo dei diversi segnali d'errore nell'aggiornamento delle rappresentazioni — se la struttura di precisione determina quale configurazione del modello generativo è resa dominante in ogni momento —, allora i sistemi neuromodulatori che regolano la precisione determinano non solo *cosa* viene rappresentato ma *come* viene rappresentato: qual è la qualità dell'esperienza, quanto è vivida, quanto è fluida, quanto è unitaria o frammentata. Le alterazioni della coscienza indotte da sostanze psicotrope che modulano i sistemi neuromodulatori — psichedelici serotoninergici come LSD e psilocibina, antidopaminergici come gli antipsicotici atipici, ansiolitici

noradrenergici come le benzodiazepine — possono essere interpretate nel framework come alterazioni della struttura di precisione del sistema predittivo che producono configurazioni del modello generativo diverse da quelle ordinarie.

Gli psichedelici serotoninergici producono fenomeni percettivi caratteristici: allucinazioni visive complesse, sinestesie intermodali, senso di profonda unità con l'ambiente circostante, dissoluzione dei confini del sé, intensificazione della vivacità delle percezioni. Nel framework predittivo, questi effetti possono essere interpretati come conseguenze di un'alterazione specifica della struttura di precisione: l'agonismo serotoninergico ai recettori 5-HT_{2A} riduce la precisione delle predizioni top-down rispetto agli errori di predizione bottom-up, producendo un sistema più «guidato dai dati» e meno «guidato dal modello» rispetto allo stato ordinario. In termini tecnici, se indichiamo con α il rapporto tra la precisione top-down e la precisione bottom-up:

$$\alpha = \Pi_{\text{topdown}} / \Pi_{\text{bottomup}}$$

lo stato psichedelico corrisponde a una riduzione di α rispetto al valore ordinario. Quando α diminuisce, i segnali sensoriali bottom-up ricevono peso relativo maggiore rispetto alle predizioni top-down: il sistema è meno in grado di sopprimere i dettagli sensoriali che il modello ordinariamente tratta come irrilevanti, producendo la vivacità e la novità percettiva caratteristiche dell'esperienza psichedelica. La dissoluzione dei confini del sé corrisponde a un'alterazione specifica della componente del modello di sé: quando la precisione del modello del confine corpo-ambiente si riduce, il segnale che distingue «questa sensazione è del mio corpo» da «questa sensazione è dell'ambiente» perde dominanza, e la struttura del modello di sé si destabilizza fenomenicamente.

VIII.7 — Il problema difficile come problema di traduzione irriducibile

La posizione sviluppata in questo capitolo può essere sintetizzata come segue: il framework predittivo offre una teoria computazionalmente precisa dei *correlati funzionali* della coscienza fenomenica — della struttura computazionale che corrisponde, nei sistemi biologici che la implementano, all'esperienza soggettiva. Questa teoria è genuinamente esplicativa nel senso che identifica i meccanismi computazionali responsabili delle caratteristiche strutturali dell'esperienza — la sua temporalità estesa, la sua unità, la sua prospettività, la sua alterabilità da parte di sostanze neuromodulatorie. Ma non è, e non pretende di essere, una riduzione dell'esperienza soggettiva ai processi computazionali: rimane aperto il problema di perché quei processi computazionali siano accompagnati da esperienza soggettiva, problema che si può formulare come il problema della traduzione irriducibile tra il vocabolario funzionale e il vocabolario fenomenologico.

Questo residuo irriducibile non è una lacuna del framework predittivo specificamente: è una lacuna di qualsiasi teoria funzionalista della coscienza, e riflette l'asimmetria profonda tra la descrizione in terza persona dei processi fisici e la descrizione in prima persona dell'esperienza soggettiva. La risposta che il framework predittivo offre al problema difficile — nella misura in cui offre una risposta — è che il problema è meno drammatico di quanto sembri una volta che si separano chiaramente le questioni funzionali dall'esperienza soggettiva: una volta che si sa con precisione quali sono i meccanismi funzionali della coscienza — la struttura del modello generativo di sé, la dinamica dell'inferenza sulle traiettorie, il ruolo della precisione neuromodulatoria — il residuo fenomenico appare non come una proprietà causalmente efficace aggiuntiva ma come un aspetto irriducibile della prospettiva in prima persona su processi che sono già completamente descritti in terza persona. Il problema difficile diventa il problema della doppia descrizione — perché la stessa realtà appare diversa dalla prospettiva in terza persona e dalla prospettiva in prima persona — piuttosto che il problema di una causa aggiuntiva misteriosamente incorporata nei processi fisici.

Questa posizione non coincide con il panpsichismo, né con il dualismo di proprietà, né con il fisicalismo eliminativista: è una forma di fisicalismo non-riduttivo che riconosce l'irriducibilità della prospettiva in prima persona senza attribuirle efficacia causale distinta dai processi computazionali che la implementano. La sua coerenza interna è quella di qualsiasi posizione non-riduttiva — fondata sull'asimmetria della prospettiva piuttosto che sull'asimmetria ontologica — e i suoi impegni metafisici sono minimali rispetto all'evidenza empirica e computazionale che il framework predittivo porta a sostegno.

IX — Sintesi: la teleosemantica computazionale sincronica

IX.1 — Formulazione sistematica della teoria

Il percorso sviluppato nei capitoli precedenti converge verso una posizione teorica definita che merita ora una formulazione sistematica e compatta, separando le tesi nucleari dagli argomenti di supporto e dalle implicazioni derivate. Si propone di denominare questa posizione *teleosemantica computazionale sincronica* — una denominazione composita che riflette i tre fattori teorici che la caratterizzano distintivamente rispetto alle teorie precedenti.

Il termine *teleosemantica* indica la continuità con la tradizione inaugurata da Millikan: il contenuto semantico è determinato da funzioni proprie, cioè da funzioni che un sistema ha in virtù di una storia di selezione — non necessariamente filogenetica nel senso della selezione naturale darwiniana, ma di qualsiasi processo di selezione che preservi selettivamente le configurazioni del sistema

che producono gli output adattativi. La componente teleologica è irrinunciabile perché fornisce la normatività che il contenuto richiede: che un sistema *rappresenti* F piuttosto che G è una questione di quale stato dovrebbe produrre date determinate condizioni, non di quale stato produce effettivamente in ogni occasione.

Il termine *computazionale* indica la specificazione del meccanismo attraverso il quale le funzioni proprie si realizzano: il framework della codifica predittiva gerarchica fornisce un'implementazione computazionale precisa del processo attraverso il quale i sistemi biologici acquisiscono e utilizzano rappresentazioni semantiche. Questa specificazione è cruciale perché risolve i problemi di indeterminazione che affliggono le teorie teleosemantiche classiche — il problema del livello di descrizione appropriato per le funzioni proprie, il problema della normatività in assenza di selezione filogenetica, il problema del contenuto astratto.

Il termine *sincronica* indica il punto teorico più innovativo e controverso: la normatività semantica è fondata sulla struttura sincronica del modello generativo nel suo stato attuale, non sulla storia diacronica della selezione che ha prodotto quel modello. Questa svolta sincronica consente di estendere la teoria al contenuto astratto e concettuale — che non ha una storia di selezione naturalizzabile senza circolarità — e di rendere conto della flessibilità semantica che caratterizza il linguaggio umano e il pensiero teorico.

Le sei tesi nucleari della teoria possono essere formulate come segue.

Tesi I — Tesi della struttura triadica del contenuto. Il contenuto semantico di uno stato cognitivo è determinato dalla tripla ($\mathbf{Q}^{(1)}$, $\mathbf{g}^{(1)}$, $\mathbf{\Pi}^{(1)}$): la distribuzione a posteriori approssimata sulle cause nascoste, la funzione generativa che mappa quelle cause sulle previsioni degli input, e la matrice di precisione che pondera la rilevanza relativa degli errori di predizione a ciascun livello gerarchico. Nessuno dei tre elementi è sufficiente da solo a determinare il contenuto: la tripla è irriducibile alle sue componenti separate.

Tesi II — Tesi dell'ancoraggio causale strutturale. L'intenzionalità degli stati cognitivi in un sistema predittivo gerarchico è non-derivativa: il contenuto è determinato dalle proprietà intrinseche del sistema — dalla struttura del modello generativo e dalla sua storia di formazione causale — senza dipendere dall'interpretazione di agenti esterni. L'ancoraggio causale strutturale soddisfa le condizioni genealogica, predittiva e controfattuale definite nel Capitolo VII.

Tesi III — Tesi della dissoluzione del problema della disgiunzione. Il problema della disgiunzione non si pone per la teleosemantica computazionale sincronica perché il contenuto

semantico non è identificato con la proprietà disgiunta che causa gli stati in tutte le occasioni, ma con la proprietà che il modello generativo ha funzione di individuare — la proprietà rispetto alla quale gli stati erronei sono *errori*, non casi alternativi di successo. Il sistema di precisione fornisce il meccanismo computazionale della correzione dell'errore che distingue la misfunzione dalla funzione normale.

Tesi IV — Tesi del radicamento nel mondo. Il significato degli stati cognitivi è determinato in parte da fattori esterni al sistema — dalla struttura dell'ambiente in cui il sistema è situato e ha acquisito il proprio modello generativo. Questa tesi è una forma di esternalismo semantico compatibile con la struttura computazionale del framework predittivo: l'esternalismo non è un'assunzione metafisica aggiuntiva ma una conseguenza della teoria dell'ancoraggio causale strutturale, che lega il contenuto alla storia di interazione con l'ambiente.

Tesi V — Tesi della normatività sincronica. La normatività semantica — il fatto che gli stati cognitivi possono essere corretti o erronei — è fondata sulla struttura del modello generativo nel suo stato attuale: la funzione propria di uno stato al livello 1 è individuare le cause nascoste $s^{(1)}$ tali che $g^{(1)}(s^{(1)})$ minimizzi l'energia libera rispetto alle osservazioni al livello 1-1. Questa funzione propria è definita sincronicamente — in relazione alla struttura attuale del modello — senza richiedere un'analisi della storia di selezione che ha prodotto il modello.

Tesi VI — Tesi della scala gerarchica del contenuto. I diversi livelli gerarchici del sistema predittivo costituiscono livelli di contenuto semantico qualitativamente distinti — dal contenuto percettivo agli stati affettivi, al contenuto concettuale astratto — che sono integrati in un unico sistema di inferenza ma hanno caratteristiche semantiche diverse in virtù della diversa struttura delle funzioni generative ai rispettivi livelli. Questa struttura scalare del contenuto è la base computazionale della ricchezza semantica del pensiero umano e della sua capacità di muoversi tra domini concreti e astratti.

IX.2 — Implicazioni per la causalità mentale

Il problema della causalità mentale è uno dei punti di maggiore tensione nella filosofia della mente contemporanea. La formulazione standard del problema — elaborata sistematicamente da Kim a partire dagli anni ottanta — parte dalla distinzione tra proprietà fisiche e proprietà mentali e dall'assunzione del determinismo fisico: se ogni evento fisico ha una causa fisica sufficiente, come possono le proprietà mentali esercitare una causalità genuina sui processi fisici? Non sembrano esserci due strade: o le proprietà mentali sono identiche a proprietà fisiche (la soluzione riduzionista che elimina il problema riducendo il mente al fisico), oppure le proprietà mentali sono causalmente inerti — *epifenomeni* dei processi fisici che accadono «in parallelo» senza

contribuire causalmente ai loro effetti. Nessuna delle due opzioni è soddisfacente: la prima elimina la mente, la seconda la rende irrilevante.

Il framework della teleosemantica computazionale sincronica offre una risposta al problema che non coincide né con il riduzionismo né con l'epifenomenismo. La risposta si articola in tre passi. Il primo passo è l'identificazione delle proprietà mentali — del contenuto semantico degli stati cognitivi — con la struttura del modello generativo: dire che un sistema crede che P equivale a dire che il modello generativo del sistema è configurato in un modo tale che P è il contenuto della rappresentazione al livello gerarchico appropriato. Questa identificazione non è un'identità tipo tra proprietà mentali e proprietà fisiche nel senso delle teorie dell'identità classiche: non richiede che ci sia un tipo fisico unico che realizza la credenza che P in tutti i possibili sistemi cognitivi. È invece un'identità istanza-per-istanza tra lo stato mentale specifico di un sistema specifico e la configurazione specifica del suo modello generativo.

Il secondo passo è il riconoscimento che la struttura del modello generativo ha efficacia causale sui processi fisici del sistema: i parametri del modello θ determinano quali predizioni sono generate, quali errori di predizione vengono calcolati, come vengono ponderati dalla struttura di precisione, come si propagano attraverso la gerarchia, e quali azioni vengono selezionate. Questa efficacia causale è completamente descritta nei termini fisici dell'implementazione neurale — è causalmente realizzata da specifici pattern di connessione sinaptica e di attività neurale — ma non è riducibile a quegli strati di descrizione senza perdita di informazione esplicativa: la struttura del modello generativo è il livello di descrizione al quale i pattern causali rilevanti diventano visibili e predicibili.

Il terzo passo è il riconoscimento che le proprietà semantiche degli stati cognitivi — il loro contenuto — supervengono sulla struttura del modello generativo nel senso fisicamente coerente della sopravvenienza: non vi può essere variazione nel contenuto semantico senza variazione nel modello generativo, e non vi può essere variazione nel modello generativo senza variazione nella struttura fisica che lo implementa. Questa catena di sopravvenienza garantisce che la causalità mentale — la causalità del contenuto semantico sulle azioni — sia causalmente ancorata alla struttura fisica del sistema senza essere identica a proprietà fisiche a grana fine. Le proprietà semantiche sono causalmente efficaci *attraverso* la loro realizzazione fisica, non *oltre* essa né *invece* di essa.

Questa posizione è una forma di non-riduzionismo compatibilista: il contenuto semantico è causalmente efficace, la sua efficacia causale è mediata dalla struttura fisica, ma la descrizione al livello del contenuto semantico è irriducibile alla descrizione al livello fisico nel senso che cattura pattern causali che la descrizione fisica

a grana fine non individua. La risposta al dilemma di Kim è che la causazione del contenuto semantico non è una causa separata dalla causazione fisica: è la descrizione della stessa causazione fisica a un livello di astrazione appropriato che ne individua la struttura semanticamente rilevante.

IX.3 — Previsioni neuroscientifiche della teoria

Una caratteristica distintiva delle teorie scientificamente serie rispetto alle speculazioni filosofiche è la capacità di generare previsioni empiricamente testabili che possano discriminare la teoria dalle sue alternative. La teleosemantica computazionale sincronica, nella misura in cui si articola in connessione con il framework della codifica predittiva, genera un insieme di previsioni neuroscientifiche che si prestano a test sperimentali con le metodologie disponibili — in particolare con la neuroimaging funzionale, l'elettrofisiologia multiunitaria, la stimolazione cerebrale e lo studio di pazienti con lesioni o disturbi cognitivi selettivi.

La prima previsione riguarda la struttura gerarchica della specializzazione semantica. Se il contenuto semantico è organizzato gerarchicamente secondo il framework predittivo, le rappresentazioni semantiche dei concetti percettivi concreti dovrebbero essere localizzate o almeno ancorate preferenzialmente nelle aree corticali più vicine alle aree sensoriali primarie — aree temporali inferiori per i concetti visivi, aree temporali superiori per i concetti uditivi, e via dicendo — mentre le rappresentazioni di concetti astratti e relazionali dovrebbero essere localizzate o ancorate preferenzialmente nelle aree corticali di ordine superiore che ricevono convergenza di input da più modalità sensoriali — la corteccia prefrontale laterale, la corteccia temporale anteriore, il cingolato anteriore. Questa previsione è in accordo con l'evidenza accumulata negli ultimi vent'anni di neuroimaging semantico, ma la teoria predittiva consente una specificazione più precisa: la distribuzione delle attivazioni per concetti astratti versus concreti dovrebbe riflettere la struttura gerarchica del modello generativo, con i concetti astratti che mostrano maggiore attivazione nelle aree di ordine superiore non perché siano lì *immagazzinati* ma perché i modelli che li computano risiedono lì.

La seconda previsione riguarda il ruolo della precisione nell'elaborazione semantica. Se la struttura di precisione — codificata dai sistemi neuromodulatori — determina il peso relativo dei diversi livelli gerarchici nell'elaborazione semantica, allora le manipolazioni farmacologiche che alterano la neuromodulazione dovrebbero produrre alterazioni sistematiche e prevedibili nell'elaborazione semantica. Specificamente, la deplezione di acetilcolina — che riduce la precisione degli input sensoriali rispetto alle predizioni top-down — dovrebbe produrre un sistema più orientato verso le rappresentazioni a priori e meno sensibile ai dettagli semantici contestuali: un sistema che mostra

maggior influenza del contesto atteso sulle inferenze semantiche e minore capacità di aggiornare la propria comprensione semantica su evidenza contrastante. Questo pattern è effettivamente osservato nei disturbi della cognizione semantica associati alla disfunzione colinergica, ma il framework fornisce una spiegazione computazionale precisa che va oltre la semplice correlazione neurobiologica.

La terza previsione riguarda la struttura degli errori semantici in condizioni di carico cognitivo. Se il contenuto semantico è determinato dalla struttura del modello generativo e dalle relative strutture di precisione, allora gli errori semantici prodotti in condizioni di carico cognitivo — quando le risorse computazionali disponibili per minimizzare l'energia libera sono ridotte — dovrebbero avere una struttura non casuale ma sistematicamente orientata verso le predizioni a priori del modello generativo. Gli errori semantici in condizioni di fatica o stress cognitivo dovrebbero essere bias verso rappresentazioni prototipiche, verso associazioni semantiche ad alta frequenza, verso interpretazioni contestualmente attese — non errori casuali ma errori che riflettono la struttura del modello generativo operante con risorse di aggiornamento ridotte.

IX.4 — Confronto con la teoria dello spazio di lavoro globale e con la teoria dell'informazione integrata

Due teorie principali della coscienza dominano il dibattito neuroscientifico contemporaneo: la teoria dello Spazio di Lavoro Globale (GWT) di Baars e Dehaene, e la teoria dell'Informazione Integrata (IIT) di Tononi. La teleosemantica computazionale sincronica ha relazioni specifiche con ciascuna di esse che meritano di essere analizzate per situare la proposta nel panorama teorico attuale.

La GWT propone che la coscienza corrisponda alla diffusione globale dell'informazione attraverso uno «spazio di lavoro» corticale distribuito che connette elaborazioni specializzate locali: uno stimolo diventa cosciente quando viene «trasmesso» dallo spazio di lavoro globale a una rete distribuita di aree corticali, producendo un'attivazione sincronizzata e amplificata. L'evidenza sperimentale per questa teoria è considerevole, in particolare gli studi su mascheramento retroattivo, accesso all'attenzione e neuroimaging di stimoli sopra e sotto-liminali. La GWT, tuttavia, non fornisce una teoria semantica: descrive le condizioni di accessibilità cosciente dell'informazione senza specificare cosa determina il contenuto semantico dell'informazione diffusa. La teleosemantica computazionale sincronica e la GWT sono dunque parzialmente ortogonali: la prima risponde alla domanda «che cosa è il contenuto semantico di uno stato cognitivo?», la seconda risponde alla domanda «quali stati cognitivi diventano coscientemente accessibili?». Una teoria completa della cognizione semantica cosciente richiede entrambe le risposte, e il framework

predittivo può in linea di principio incorporare una versione computazionale della GWT — identificando lo «spazio di lavoro globale» con i livelli superiori della gerarchia predittiva che generano predizioni con ampio range di influenza sui livelli inferiori.

La IIT propone che la coscienza sia identica all'informazione integrata misurata dalla quantità Φ (phi): la coscienza è reale nella misura in cui un sistema genera più informazione come totalità di quanta ne generino le sue parti considerate separatamente. La IIT ha l'ambizione di fornire una misura matematicamente precisa della coscienza applicabile in linea di principio a qualsiasi sistema fisico, il che la rende estremamente generale ma anche controversa: ha implicazioni panpsichiste che molti ricercatori trovano difficili da accettare, e la computazione di Φ è computazionalmente intrattabile per sistemi reali di dimensioni cerebrali. Il confronto con la teleosemantica computazionale sincronica è interessante perché le due teorie hanno diverse strutture normative: la IIT fonda la coscienza sull'integrazione dell'informazione come proprietà strutturale del sistema, la teleosemantica la fonda sul contenuto semantico come proprietà relazionale del sistema rispetto all'ambiente. Un sistema altamente integrato nel senso della IIT non ha necessariamente un ricco contenuto semantico se non è adeguatamente radicato in un ambiente strutturato; conversamente, un sistema semanticamente ricco può avere un Φ relativamente basso se la sua integrazione è distribuita in modo da non soddisfare i criteri della IIT. Le due teorie non sono rivali nel senso diretto: affrontano aspetti diversi della cognizione cosciente, e la loro integrazione potrebbe fornire una teoria più completa di quanto ciascuna delle due sia in grado di fornire separatamente.

IX.5 — La teleosemantica computazionale sincronica nel panorama del naturalismo semantico

Il naturalismo semantico ha attraversato tre fasi principali nella seconda metà del Novecento, ognuna delle quali ha aperto nuove possibilità teoriche ma ha lasciato irrisolti problemi specifici che hanno motivato lo sviluppo della fase successiva. La prima fase — quella dell'informazione causal-nomologica (Dretske, Fodor) — ha stabilito che il contenuto semantico può essere naturalizzato attraverso relazioni causali tra stati del sistema e proprietà del mondo, ma ha incontrato il problema della disgiunzione e il problema del contenuto normativo. La seconda fase — quella della teleosemantica evoluzionistica (Millikan, Papineau, Neander) — ha risolto il problema normativo attraverso l'appello alla selezione naturale, ma ha incontrato il problema del contenuto astratto e dell'estensibilità diacronica. La terza fase — quella della teleosemantica computazionale sincronica proposta in questa monografia — tenta di risolvere i problemi lasciati aperti dalla seconda fase attraverso la specificazione computazionale del meccanismo di assegnazione della funzione propria e la svolta

sincronica che radica la normatività nella struttura attuale del modello generativo piuttosto che nella sua storia di selezione.

Questa collocazione storica consente di vedere la proposta come una risposta agli sviluppi della filosofia della mente e della scienza cognitiva piuttosto che come una proposta isolata: essa dialoga con le teorie precedenti, conserva quanto di esse è difendibile, e tenta di superare i problemi specifici che le hanno rese vulnerabili alle obiezioni. La continuità con Millikan è la più profonda: la normatività del contenuto richiede funzioni proprie, e le funzioni proprie richiedono una storia selettiva. La discontinuità con Millikan è la svolta sincronica: la storia selettiva rilevante non è la storia filogenetica della selezione naturale ma la storia ontogenetica dell'apprendimento percettivo e concettuale, e la normatività che ne risulta è fondata sulla struttura sincronica del modello generativo attuale piuttosto che su un'analisi teleologica della forma degli organi di rappresentazione.

X — Il linguaggio del pensiero rivisitato: sistemi ibridi e contenuto proposizionale

X.1 — La sfida del linguaggio del pensiero

La teoria del linguaggio del pensiero, elaborata da Fodor nella sua formulazione più ambiziosa in *The Language of Thought* (1975) e difesa con tenacia nelle revisioni successive fino a *LOT 2* (2008), postula che il pensiero si svolga in un medium rappresentazionale con struttura compositiva analoga alla struttura del linguaggio naturale: vi è un vocabolario di simboli primitivi, regole di combinazione sintattica che generano strutture complesse, e proprietà semantiche che si proiettano composizionalmente dalla semantica dei primitivi alla semantica delle strutture complesse. Questa ipotesi consente di rendere conto di due proprietà fondamentali del pensiero umano che qualsiasi teoria cognitiva deve spiegare: la sistematicità — il fatto che chiunque possa pensare «il gatto è sul tappeto» può anche pensare «il tappeto è sotto il gatto» — e la produttività — il fatto che il numero di pensieri distinti che un agente cognitivo può formare è in linea di principio illimitato pur disponendo di un numero finito di primitivi semantici.

La LOT è strettamente connessa al classicismo computazionale — all'ipotesi che la cognizione sia computazione su strutture simboliche — e costituisce uno degli obiettivi polemici privilegiati dei difensori delle reti neurali e dei modelli sub-simbolici. La codifica predittiva gerarchica è un modello sub-simbolico nel senso che le rappresentazioni ai livelli della gerarchia non hanno la struttura discreta e combinatoria dei simboli fodoriani: sono distribuzioni di probabilità su spazi continui di variabili latenti, le cui strutture compositive non sono manifeste nella struttura della

singola distribuzione ma emergono attraverso la struttura della gerarchia delle funzioni generative.

La domanda che si pone nell'ambito della teleosemantica computazionale sincronica è se il framework predittivo sia in grado di rendere conto della sistematicità e della produttività del pensiero senza dover presupporre un livello di rappresentazione con struttura LOT. La risposta che si difende qui non è né l'affermazione incondizionata né la negazione incondizionata: si propone piuttosto una visione ibrida che distingue il pensiero percettivo-contestuale — per il quale le rappresentazioni distribuite del framework predittivo sono sufficienti — dal pensiero proposizionale esplicito — per il quale una struttura simile alla LOT emerge come livello funzionale superiore realizzato nelle aree prefrontali e nel linguaggio interiore.

X.2 — Sistematicità e struttura compositiva nei modelli predittivi

L'argomento di Fodor per la LOT a partire dalla sistematicità è considerato il più forte: se la sistematicità del pensiero fosse una proprietà contingente — se qualcuno potesse pensare «il gatto è sul tappeto» senza necessariamente poter pensare «il tappeto è sotto il gatto» — ciò richiederebbe una spiegazione specifica per ogni coppia di pensieri sistematicamente correlati, una spiegazione che si moltiplica indefinitamente. La spiegazione più parsimoniosa della sistematicità è che i pensieri condividono componenti semantiche strutturate compositivamente: «il gatto», «il tappeto», «è sul» sono componenti che si ricombinano sistematicamente, e chiunque padroneggi le componenti padroneggia tutte le loro combinazioni grammaticalmente lecite.

Come risponde il framework predittivo a questo argomento? La risposta preferita dai sostenitori dei modelli sub-simbolici — che la sistematicità è una proprietà statistica emergente dalle reti addestrate su input sistematici — è stata giudicata insufficiente da Fodor, che ha argomentato che l'emergenza statistica non garantisce la necessità logica della sistematicità. Tuttavia, il framework predittivo gerarchico offre una risorsa aggiuntiva che i modelli connessionisti classici non avevano: la struttura gerarchica delle funzioni generative. Se le funzioni generative ai livelli superiori della gerarchia codificano relazioni strutturali tra rappresentazioni — relazioni di tipo «X è sul Y», «X appartiene a Y», «X causa Y» — nella loro struttura parametrica, allora la combinabilità delle rappresentazioni è garantita non dall'emergenza statistica ma dalla struttura delle funzioni generative stesse.

Formalmente, si può proporre che le funzioni generative ai livelli superiori abbiano la struttura

$$\mathbf{g}^{(l)}(\mathbf{s}^{(l)}) = \mathbf{R}(\mathbf{s}_{\text{rel}}^{(l)}, \mathbf{s}_{\text{arg1}}^{(l)}, \mathbf{s}_{\text{arg2}}^{(l)})$$

dove $s_rel^{(1)}$ è la rappresentazione della relazione («è sul», «appartiene a», «causa»), $s_arg1^{(1)}$ e $s_arg2^{(1)}$ sono le rappresentazioni degli argomenti, e R è una funzione che combina questi elementi in una predizione delle rappresentazioni al livello inferiore. Una tale struttura garantisce che chiunque abbia una rappresentazione della relazione «è sul» e una rappresentazione di «gatto» e «tappeto» possa generare le predizioni sia per «il gatto è sul tappeto» sia per «il tappeto è sotto il gatto» — con l'inversione degli argomenti e la corrispondente variazione nella relazione. La sistematicità è così una proprietà strutturale delle funzioni generative ai livelli superiori della gerarchia, non una proprietà emergente statisticamente.

X.3 — La teoria allostatica del contenuto affettivo

Una delle estensioni più significative e feconde del framework predittivo riguarda il contenuto degli stati affettivi — emozioni, stati d'umore, affetti di valenza — che la tradizione filosofica ha spesso trattato come intrinsecamente refrattari alla naturalizzazione semantica. La proposta di Corcoran e Bhatt, sviluppata in dialogo con il framework frinstonianiano, identifica gli stati affettivi con inferenze predittive relative agli stati interni del corpo piuttosto che agli stati esterni dell'ambiente: gli stati affettivi sono predizioni sull'omeostasi e l'allostasi del sistema biologico — sulla deviazione degli stati fisiologici interni dai loro valori di riferimento e sulla previsione di future deviazioni.

L'allostasi, a differenza dell'omeostasi, è la modulazione anticipatoria dei parametri fisiologici in risposta a previsioni di domanda futura piuttosto che a deviazioni attuali dai valori di riferimento: il sistema cardiovascolare e il sistema neuroendocrino si attivano in anticipazione dello stress fisico o psicologico, preparandosi a un'alterazione che non si è ancora verificata. In termini predittivi, l'allostasi è la generazione di predizioni degli stati fisiologici futuri e la modulazione degli stati presenti per minimizzare l'errore di predizione sui futuri stati fisiologici. Gli stati affettivi, in questa prospettiva, sono il correlato esperienziale — la componente del modello di sé — delle predizioni allostatiche: la paura è la predizione di stati fisiologici di minaccia, il desiderio è la predizione di stati fisiologici di carenza, la gioia è la disconferma di predizioni negative o la conferma di predizioni positive di stati fisiologici benigne.

Il contenuto semantico degli stati affettivi è determinato dalla tripla triadica $Q^{(1)}$, $g^{(1)}$, $\Pi^{(1)}$ applicata agli stati interocettivi: la distribuzione a posteriori approssimata sugli stati fisiologici interni, la funzione generativa che mappa quegli stati sulle previsioni degli input interocettivi, e la precisione con cui quegli input vengono ponderati rispetto alle predizioni top-down del modello allostatico. La valenza affettiva — il carattere di piacevolezza o spiacevolezza di uno stato — è determinata dal segno dell'errore di predizione allostatico: gli stati fisiologici che confermano predizioni positive o

disconfermano predizioni negative sono vissuti come positivi, gli stati che confermano predizioni negative o disconfermano predizioni positive sono vissuti come negativi. La dimensione arousal — l'intensità dell'attivazione affettiva — è determinata dalla magnitudine dell'errore di predizione allostasico: quanto più la deviazione dei valori fisiologici dalle predizioni del modello allostasico è ampia, tanto più intensa è l'esperienza affettiva.

X.4 — Il problema del contenuto proposizionale e la sua soluzione

Il contenuto proposizionale — il contenuto degli stati cognitivi che si esprime con proposizioni — ha struttura valutativa in base a valori di verità: la credenza che «p» è vera se e solo se è il caso che p. La struttura vero-condizionale del contenuto proposizionale ha generato la tradizione semantica delle condizioni di verità, da Frege attraverso Tarski, Davidson e Lewis, e ha costituito uno degli ostacoli più seri per le teorie naturaliste del contenuto: come si naturalizzano le condizioni di verità?

Nel framework della teleosemantica computazionale sincronica, le condizioni di verità di un contenuto proposizionale sono naturalizzate attraverso la struttura della distribuzione a posteriori $Q^{(1)}$ e della funzione generativa $g^{(1)}$ al livello gerarchico appropriato. La proposizione «il gatto è sul tappeto» è vera, nel framework, se e solo se la configurazione dello stato nascosto $s^{(1)}$ che massimizza $Q^{(1)}(s^{(1)}; \mu^{(1)})$ è una configurazione che include la relazione spaziale di sovrapposizione tra le rappresentazioni del gatto e del tappeto — cioè se l'inferenza predittiva converge su quella configurazione dello stato nascosto come la spiegazione più probabile degli input sensoriali disponibili. Le condizioni di verità sono le condizioni sullo stato del mondo alle quali la distribuzione a posteriori converge su quella configurazione piuttosto che su alternative: la proposizione è vera quando il mondo è configurato in modo che la minimizzazione dell'energia libera rispetto alle osservazioni disponibili porta il sistema a rappresentare quella configurazione come la più probabile.

Questa naturalizzazione delle condizioni di verità non elimina la verità come proprietà semantica irriducibile: mantiene la distinzione tra la credenza di un agente e la sua verità — un agente può credere che p (cioè il suo modello generativo può convergere su quella configurazione) anche quando p è falsa (cioè anche quando quella configurazione non corrisponde allo stato effettivo del mondo). Ma ancora, la distinzione tra credenza e verità è naturalizzata: la credenza è la configurazione del modello generativo, la verità è la corrispondenza tra quella configurazione e lo stato effettivo del mondo — una corrispondenza che si esprime, nel lungo periodo, attraverso la pressione selettiva della minimizzazione dell'energia libera verso configurazioni del modello che siano adeguate all'ambiente.

X.5 — Composizionalità e apprendimento

La composizionalità del significato — il principio Frege-Carnap secondo cui il significato di un'espressione complessa è funzione del significato delle sue parti — è un vincolo fondamentale per qualsiasi teoria semantica che pretenda di essere adeguata alla ricchezza del linguaggio naturale e del pensiero proposizionale. Nel framework predittivo, la composizionalità è realizzata dalla struttura gerarchica del modello generativo: il significato di «gatto bianco sul tappeto rosso» è determinato compositivamente dal significato di «gatto», «bianco», «tappeto», «rosso» e dalla struttura delle relazioni di modificazione e locazione che le funzioni generative ai livelli superiori della gerarchia codificano.

La questione dell'apprendimento della composizionalità — come i sistemi predittivi acquisiscono funzioni generative con struttura compositiva adeguata — è una delle questioni empiriche più urgenti nel campo della cognizione computazionale. L'evidenza dagli studi su reti neurali addestrate su testi linguistici mostra che modelli sufficientemente grandi e addestrati su corpora sufficientemente vasti sviluppano spontaneamente strutture rappresentazionali con proprietà composizionali di fatto — strutture che generalizzano in modo sistematicamente compositivo a combinazioni non viste durante l'addestramento. Questo suggerisce che la composizionalità è una proprietà che emerge dall'ottimizzazione su dati composizionali, non una struttura imposta dall'esterno al sistema di apprendimento.

Nel framework della codifica predittiva, questa emergenza della composizionalità può essere spiegata come segue: un sistema che apprende un modello generativo da un ambiente linguistico in cui le espressioni composte sistematicamente ricorrono con regolarità prevedibile ricaverà funzioni generative con struttura compositiva perché la struttura compositiva è *la struttura causale* dell'ambiente linguistico — è il meccanismo che genera la regolarità prevedibile delle combinazioni semantiche. Un modello che apprende quella struttura causale attraverso la minimizzazione dell'energia libera apprende necessariamente la composizionalità.

XI — Evidenze neuroscientifiche, semantica distribuzionale e pragmatica

XI.1 — L'evidenza neuroimaging per la gerarchia predittiva

Le previsioni del framework predittivo gerarchico hanno ricevuto supporto empirico considerevole dalla neuroimaging funzionale degli ultimi quindici anni. In particolare, tre linee di evidenza convergono a sostegno della struttura gerarchica del processamento predittivo nella corteccia cerebrale.

La prima linea riguarda la struttura delle risposte BOLD in fMRI a stimoli prevedibili versus imprevedibili. Se la corteccia implementa un sistema predittivo gerarchico, le risposte alle violazioni delle aspettative — agli stimoli imprevedibili dati il contesto precedente — dovrebbero essere più ampie rispetto alle risposte a stimoli prevedibili: gli errori di predizione producono segnali più prominenti rispetto alle conferme di predizione. Questo pattern — denominato *mismatch* o risposta alla soppressione della ripetizione — è stato documentato in numerosi studi di adattamento fMRI e violazione delle aspettative, in diverse modalità sensoriali e livelli gerarchici di processamento. La regressione lineare tra la sorpresa bayesiana delle transizioni linguistiche, visive e uditive e la risposta BOLD nelle rispettive aree corticali ha mostrato correlazioni significative che supportano l'interpretazione predittiva.

La seconda linea riguarda la struttura spaziale delle risposte predittive nella gerarchia corticale. Il framework predittivo prevede che i segnali di predizione (top-down) e i segnali di errore di predizione (bottom-up) abbiano distribuzioni spaziali distinte nella struttura laminare della corteccia: le predizioni top-down sarebbero trasmesse principalmente attraverso le connessioni cortico-corticali discendenti che terminano negli strati superficiali delle aree di ordine inferiore (strato I-II), mentre gli errori di predizione sarebbero trasmessi attraverso le connessioni ascendenti che originano dagli strati profondi (strato V-VI) e terminano negli strati granulari (strato IV) delle aree di ordine superiore. Questa previsione laminare è stata confermata in studi di registrazione multiunitaria in primati non umani che mostrano il pattern atteso di soppressione degli errori di predizione negli strati granulari delle aree di ordine superiore in risposta a stimoli prevedibili.

La terza linea riguarda il ruolo della precisione nell'attenzione selettiva. Se l'attenzione è codificata come modulazione della precisione — come proposto da Friston e Spratling —, le manipolazioni dell'attenzione dovrebbero produrre variazioni nella magnitudine degli errori di predizione ponderati dalla precisione senza necessariamente alterare i valori degli errori non ponderati. L'evidenza da studi di attenzione selettiva con neuroimaging mostra che l'attenzione amplifica le risposte corticali agli stimoli rilevanti in modo congruente con un'amplificazione della precisione degli errori di predizione a quei livelli gerarchici che elaborano le caratteristiche degli stimoli rilevanti.

XI.2 — Predizione linguistica e contenuto semantico

L'elaborazione linguistica è il dominio in cui le previsioni del framework predittivo hanno ricevuto il supporto empirico più robusto e metodologicamente sofisticato. Gli studi su effetti di facilitazione semantica e pragmatica con misure elettrofisiologiche — in particolare la componente negativa N400 dell'ERP, sensibile alla difficoltà di integrazione semantica — mostrano in modo

consistente che le parole semanticamente congruenti con il contesto precedente producono risposte N400 ridotte rispetto a parole semanticamente incongruenti, e che questa riduzione è graduata in modo continuo con la prevedibilità semantica della parola nel contesto. Il gradiente di N400 in funzione della sorpresa semantica è uno dei risultati più replicati della neurolinguistica cognitiva e costituisce un'evidenza diretta per il processamento predittivo della semantica nel linguaggio.

Nella cornice teorica della teleosemantica computazionale sincronica, la componente N400 riflette la magnitudine dell'errore di predizione semantica al livello della gerarchia predittiva che elabora il contenuto semantico lessicale: le parole ad alta prevedibilità producono errori di predizione piccoli e dunque risposte N400 ridotte, le parole a bassa prevedibilità producono errori di predizione ampi e risposte N400 ampliate. La struttura fine della distribuzione di N400 — la sua modulazione da parte del contesto semantico a vari livelli di distanza sintattica, da parte del contesto pragmatico incluso il contesto extralinguistico, da parte dello stato di conoscenza attribuito all'interlocutore — è coerente con un sistema che mantiene un modello generativo multilivello del discorso in corso, aggiornando continuamente le predizioni semantiche in risposta all'evidenza accumulata a molteplici scale temporali e contestuali.

La semantica distribuzionale — l'approccio che rappresenta il significato lessicale attraverso vettori in spazi di alta dimensione ricavati dalla co-occorrenza dei termini in corpora linguistici — ha un'interpretazione naturale nel framework predittivo: gli spazi di incorporamento semantico sono approssimazioni delle distribuzioni di probabilità condizionale che il modello generativo mantiene sugli stati nascosti linguistici, e la similarità coseno tra vettori di incorporamento riflette la prossimità semantica come sovrapposizione di probabilità a posteriori. I modelli linguistici di grandi dimensioni che utilizzano architetture transformer possono essere visti come implementazioni computazionali approssimative di sistemi predittivi gerarchici addestrati sulla minimizzazione dell'errore di predizione linguistica: la loro capacità di catturare regolarità semantiche complesse è coerente con il framework predittivo, anche se la corrispondenza con l'implementazione biologica è lontana e molti aspetti dell'architettura transformer — in particolare l'attenzione come meccanismo di ponderazione — hanno analogie ma non identità con i meccanismi di precisione del framework predittivo biologicamente plausibile.

XI.3 — I large language model come sistemi predittivi non incarnati

L'avvento dei modelli linguistici di grandi dimensioni (LLM) ha aperto una nuova e controversa linea di ricerca sulle capacità semantiche dei sistemi computazionali sub-simbolici. I modelli come GPT-4, Claude e Gemini mostrano capacità impressionanti

in compiti di ragionamento semantico, analogia, inferenza pragmatica e produzione di testo coerente — capacità che sembravano richiedere, nella prospettiva fodoriana classica, una LOT esplicita con struttura compositiva. La questione filosofica fondamentale è se queste capacità costituiscano evidenza per la tesi che sistemi predittivi sub-simbolici possono in linea di principio avere contenuto semantico genuino, o se siano mere simulazioni di competenza semantica fondate su pattern statistici che imitano il comportamento senza possedere il contenuto.

La posizione della teleosemantica computazionale sincronica rispetto a questa questione è rigorosamente strutturalista: la presenza o assenza di contenuto semantico genuino in un sistema dipende dalla struttura del modello generativo e dalla sua storia di formazione causale, non dalla natura del substrato fisico. Un LLM che abbia appreso un modello generativo del linguaggio attraverso l'ottimizzazione della previsione su un corpus sufficiente possiede, in virtù di quella struttura e di quella storia, una forma di contenuto semantico determinata dalla tripla triadica ($\mathbf{Q}$, $\mathbf{g}$, $\mathbf{\Pi}$) applicata al dominio linguistico. La questione non è se gli LLM abbiano un contenuto semantico «vero» o «simulato» — questa distinzione presuppone un dualismo tra contenuto e meccanismo che la teoria rifiuta — ma quale sia la struttura e la completezza del contenuto semantico che possiedono, e in che misura differisca dal contenuto semantico di agenti biologici incarnati.

La differenza principale tra gli LLM e i sistemi cognitivi biologici nella prospettiva del framework predittivo è l'incarnazione e la chiusura corporea. I sistemi biologici hanno un modello generativo che include non solo le strutture linguistiche e concettuali ma anche le strutture sensorimotorie e interocettive che radicano il contenuto nel corpo e nell'azione. Il contenuto del concetto «rosso» per un agente biologico è determinato dalla tripla triadica che include i livelli visivi inferiori — le funzioni generative che codificano le proprietà di riflessione della luce, la struttura del sistema dei coni retinici, le predizioni sulle sequenze di esperienza visiva con oggetti rossi. Il contenuto del «token» «rosso» in un LLM è determinato esclusivamente dalla struttura delle co-occorrenze linguistiche in cui quel token appare nel corpus di addestramento — una struttura che è una proiezione del contenuto biologico nel dominio linguistico ma non ne possiede la radice percettiva. Questa differenza non rende il contenuto degli LLM semanticamente nullo: lo rende semanticamente incompleto rispetto al contenuto biologico pienamente incarnato, e potenzialmente adeguato per i domini in cui il contenuto linguistico è autosufficiente — matematica, logica formale, argomentazione filosofica — mentre è inadeguato per i domini in cui il contenuto richiede l'ancoraggio percettivo diretto.

XI.4 — Pragmatica dell'interpretazione come inferenza predittiva

La pragmatica del linguaggio — la dimensione dell'uso contestuale delle espressioni linguistiche che va oltre le condizioni di verità semantiche — ha costituito tradizionalmente un dominio separato dalla semantica, studiato con strumenti teorici distinti: le massime conversazionali di Grice, i principi della teoria della pertinenza di Sperber e Wilson, le norme della teoria degli atti linguistici di Austin e Searle. Il framework predittivo offre una cornice unificata in cui la pragmatica è continua con la semantica: entrambe sono forme di inferenza predittiva su contenuti comunicativi, distinte per il livello gerarchico al quale operano e per il tipo di modello generativo che le governa.

La comprensione pragmatica — l'inferenza dell'intenzione comunicativa dell'interlocutore oltre il contenuto semantico letterale dell'enunciato — è un'inferenza bayesiana su stati nascosti di ordine superiore: l'intenzione dell'interlocutore, le sue credenze sulle credenze del ricevente, il contesto della conversazione, le norme sociali e comunicative in vigore. Il modello generativo che governa la comprensione pragmatica include una componente di teoria della mente — un modello degli stati mentali dell'interlocutore — che genera predizioni sui comportamenti linguistici e para-linguistici di quell'interlocutore in funzione delle sue credenze, desideri e intenzioni inferite. L'inferenza pragmatica è la minimizzazione dell'energia libera rispetto a questo modello: dato l'enunciato prodotto dall'interlocutore, qual è la configurazione degli stati mentali dell'interlocutore che meglio spiega quell'enunciato nel contesto comunicativo in corso?

La rilevanza nella teoria di Sperber e Wilson può essere reinterpretata nel framework come un principio di ottimizzazione della relazione tra il costo computazionale dell'inferenza e il guadagno informativo nella minimizzazione dell'energia libera: un'inferenza è pertinente nella misura in cui produce una riduzione significativa dell'energia libera a un costo computazionale sostenibile. Le implicature conversazionali di Grice — le inferenze pragmatiche che il ricevente trae dalla violazione apparente delle massime di qualità, quantità, pertinenza e modo — sono, nel framework predittivo, errori di predizione rispetto al modello generativo delle norme comunicative cooperative che spingono il ricevente ad aggiornare il proprio modello degli stati mentali dell'interlocutore per spiegare la violazione apparente. L'implicatura «stai insinuando che sono in ritardo?» a fronte dell'enunciato «il treno delle diciassette è già partito» è l'aggiornamento del modello degli stati mentali dell'interlocutore — l'inferenza che egli intende comunicare la conseguenza che il ricevente non era sul treno — come la spiegazione più parsimoniosa della scelta di quell'enunciato in quel contesto.

XI.5 — RSA e il modello bayesiano della pragmatica

Il modello di Pragmatica a Campionamento Razionale (Rational Speech Act, RSA) di Frank e Goodman formalizza l'inferenza pragmatica come un processo bayesiano ricorsivo tra modelli di ascoltatori e parlanti razionali: un parlante razionale sceglie l'enunciato che un ascoltatore razionale interpreterebbe correttamente nel modo più efficiente, e un ascoltatore razionale aggiorna le proprie credenze sugli stati del mondo ponderando le probabilità degli enunciati che un parlante razionale avrebbe scelto per comunicare diversi stati. Questa struttura ricorsiva — che si chiude su un modello letterale di base — ha mostrato di catturare molti fenomeni pragmatici con precisione quantitativa, incluse le distribuzioni di probabilità sulle interpretazioni scalari come «alcuni» e i processi di riferimento in contesti visivi.

La connessione tra RSA e il framework predittivo è strutturalmente profonda: entrambi sono formulazioni bayesiane dell'inferenza pragmatica, e la differenza principale è il livello di astrazione a cui operano. RSA è un modello computazionale dell'algoritmo pragmatico — una teoria a livello di implementazione dell'inferenza dell'intenzione comunicativa — mentre il framework predittivo è una teoria del meccanismo di apprendimento e aggiornamento in tempo reale che implementa quell'algoritmo. La connessione suggerisce che il modello generativo dell'ascoltatore bayesiano di RSA è un'astrazione computazionale del modello generativo predittivo della gerarchia corticale che elabora il linguaggio.

Appendice A — Glossario matematico

A.1 — Notazione e convenzioni

Il glossario che segue introduce la notazione utilizzata in questa monografia e definisce formalmente i concetti matematici fondamentali del framework della codifica predittiva gerarchica. Si seguono le convenzioni standard della letteratura sull'inferenza attiva e sulla codifica predittiva, con alcune scelte notazionali specifiche introdotte per chiarezza espositiva.

Scalari, vettori, matrici. Le lettere minuscole romane indicano scalari (a , b , c); le lettere minuscole in grassetto indicano vettori colonna ($\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$); le lettere maiuscole in grassetto indicano matrici ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$). I vettori sono vettori colonna per default; i vettori riga si indicano con la notazione di trasposizione $\mathbf{a}^T$. Il prodotto scalare di due vettori $\mathbf{a}$ e $\mathbf{b}$ della stessa dimensione è $\mathbf{a}^T \mathbf{b} = \sum_i a_i b_i$.

Insiemi e spazi. $\mathbb{R}$ denota i numeri reali; $\mathbb{R}^n$ denota lo spazio euclideo n -dimensionale; $\mathbb{R}_+$ denota i numeri reali positivi; $\mathbb{N}$ denota i numeri naturali incluso lo zero. Per uno spazio vettoriale V , $\dim(V)$ denota la sua dimensione.

Distribuzioni di probabilità. $P(\cdot)$ denota distribuzioni di probabilità esatte; $Q(\cdot)$ denota distribuzioni di probabilità approssimate variazionali. $p(\cdot)$ e $q(\cdot)$ denotano le rispettive densità di probabilità quando le variabili sono continue. $E_Q[\cdot]$ denota il valore atteso rispetto alla distribuzione Q . Per una distribuzione gaussiana di media μ e matrice di covarianza Σ si scrive $Q(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu, \Sigma)$.

Divergenza KL. La divergenza di Kullback-Leibler tra due distribuzioni Q e P sullo stesso spazio è

$$D_{\text{KL}}[Q \parallel P] = E_Q[\log Q(\mathbf{s}) - \log P(\mathbf{s})] = \int q(\mathbf{s}) \log (q(\mathbf{s})/p(\mathbf{s})) d\mathbf{s}$$

La divergenza KL è sempre non-negativa: $D_{\text{KL}}[Q \parallel P] \geq 0$ con uguaglianza se e solo se $Q = P$ quasi ovunque. Non è simmetrica: $D_{\text{KL}}[Q \parallel P] \neq D_{\text{KL}}[P \parallel Q]$ in generale.

Operatori differenziali. $\partial f / \partial \mathbf{x}$ denota il gradiente di uno scalare f rispetto al vettore $\mathbf{x}$ — un vettore colonna delle derivate parziali. $\partial \mathbf{f} / \partial \mathbf{x}$ denota la matrice Jacobiana della funzione vettoriale $\mathbf{f}$ rispetto al vettore $\mathbf{x}$ — la matrice delle derivate parziali di ogni componente di $\mathbf{f}$ rispetto a ogni componente di $\mathbf{x}$. $\nabla^2 f$ denota la matrice Hessiana delle derivate seconde di f .

A.2 — Energia libera variazionale: derivazione

L'energia libera variazionale $\mathbf{F}[Q, \mathbf{o}]$ è definita come un limite superiore dell'energia libera di sorpresa $-\log P(\mathbf{o})$:

$$\mathbf{F}[Q, \mathbf{o}] = D_{\text{KL}}[Q(\mathbf{s}) \parallel P(\mathbf{s} \mid \mathbf{o})] - \log P(\mathbf{o})$$

Poiché $D_{\text{KL}}[Q \parallel P(\mathbf{s} \mid \mathbf{o})] \geq 0$, segue che $\mathbf{F}[Q, \mathbf{o}] \geq -\log P(\mathbf{o})$, con uguaglianza se e solo se $Q(\mathbf{s}) = P(\mathbf{s} \mid \mathbf{o})$. La minimizzazione di $\mathbf{F}$ rispetto a Q è dunque equivalente — al limite — all'inferenza bayesiana esatta. La formula può essere riscritta attraverso algebra elementare in tre forme equivalenti:

Prima decomposizione (complessità — accuratezza):

$$\mathbf{F}[Q, \mathbf{o}] = D_{\text{KL}}[Q(\mathbf{s}; \lambda) \parallel P(\mathbf{s})] - E_Q[\log P(\mathbf{o} \mid \mathbf{s})]$$

La prima componente — divergenza KL tra la a posteriori approssimata e la a priori — misura la complessità della distribuzione a posteriori rispetto al modello di prior. La seconda componente — log-verosimiglianza attesa — misura la bontà di adattamento del modello all'evidenza osservata. Minimizzare $\mathbf{F}$ rispetto ai parametri variazionali λ minimizza la complessità massimizzando la bontà di adattamento.

Seconda decomposizione (entropia — energia):

$$\mathbf{F}[Q, \mathbf{o}] = -H[Q(\mathbf{s}; \lambda)] + E_Q[-\log P(\mathbf{s}, \mathbf{o})]$$

dove $H[Q] = -E_Q[\log Q]$ è l'entropia di Shannon della distribuzione Q . Questa decomposizione mostra che minimizzare

$\mathbf{F}$ è equivalente a trovare il compromesso ottimale tra massimizzare l'entropia della distribuzione a posteriori (favorire distribuzioni diffuse) e massimizzare la probabilità congiunta dell'energia (favorire stati ad alta densità di probabilità congiunta $\mathbf{s}$, $\mathbf{o}$).

Terza decomposizione (KL – sorpresa attesa):

$$\mathbf{F}[\mathbf{Q}, \mathbf{o}] = D_{\text{KL}}[\mathbf{Q}(\mathbf{s}; \lambda) \parallel \mathbf{P}(\mathbf{s} \mid \mathbf{o})] + (-\log \mathbf{P}(\mathbf{o}))$$

Questa decomposizione mostra esplicitamente che $\mathbf{F}$ = divergenza dall'inferenza ottimale + sorpresa sull'osservazione. Il primo termine tende a zero man mano che $\mathbf{Q}$ si avvicina alla vera a posteriori; il secondo termine è indipendente da $\mathbf{Q}$ e rappresenta la quantità di sorpresa irriducibile data l'evidenza.

A.3 — Approssimazione di Laplace

Nella codifica predittiva gerarchica, l'approssimazione variazionale è tipicamente l'approssimazione di Laplace: $\mathbf{Q}(\mathbf{s}; \boldsymbol{\mu}) = \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, dove la covarianza $\boldsymbol{\Sigma} = (-\partial^2 \log \mathbf{P}(\mathbf{s}, \mathbf{o}) / \partial \mathbf{s}^2) |_{\mathbf{s}=\boldsymbol{\mu}}^{-1}$ è la matrice Hessiana invertita del log-posteriore valutata nel modo $\boldsymbol{\mu}$. Questa approssimazione riduce il problema variazionale alla ricerca del modo del log-posteriore — un problema di ottimizzazione anziché di integrazione — e consente aggiornamenti in tempo reale attraverso la discesa del gradiente sull'energia libera.

Nella formulazione a tempo continuo, l'equazione di aggiornamento per il parametro variazionale $\boldsymbol{\mu}^{(l)}$ al livello l è:

$$\dot{\boldsymbol{\mu}}^{(l)} = -\partial \mathbf{F} / \partial \boldsymbol{\mu}^{(l)} = (\partial \mathbf{g}^{(l)} / \partial \boldsymbol{\mu}^{(l)})^T \boldsymbol{\xi}^{(l)} - \boldsymbol{\xi}^{(l+1)}$$

dove l'errore di predizione al livello l è $\boldsymbol{\epsilon}^{(l)} = \boldsymbol{\mu}^{(l-1)} - \mathbf{g}^{(l)}(\boldsymbol{\mu}^{(l)})$ e l'errore ponderato è $\boldsymbol{\xi}^{(l)} = \boldsymbol{\Pi}^{(l)} \boldsymbol{\epsilon}^{(l)}$. La regola di aggiornamento è dunque una discesa del gradiente sull'energia libera rispetto ai parametri variazionali — un processo di inferenza in tempo reale che converge al modo della distribuzione a posteriori approssimata.

A.4 — Connessione con il filtro di Kalman

Il filtro di Kalman è il caso speciale lineare-gaussiano della codifica predittiva gerarchica. Quando le funzioni generative sono lineari — $\mathbf{g}^{(l)}(\mathbf{s}^{(l)}) = \mathbf{A}^{(l)} \mathbf{s}^{(l)}$ per una matrice $\mathbf{A}^{(l)}$ — e le distribuzioni sono gaussiane, le equazioni di aggiornamento del framework predittivo si riducono alle equazioni del filtro di Kalman esteso. Le predizioni $\mathbf{g}^{(l)}(\boldsymbol{\mu}^{(l)}) = \mathbf{A}^{(l)} \boldsymbol{\mu}^{(l)}$ sono le predizioni lineari del filtro; gli errori di predizione $\boldsymbol{\epsilon}^{(l)}$ sono le innovazioni del filtro; le matrici di precisione $\boldsymbol{\Pi}^{(l)}$ sono le matrici di guadagno di Kalman scalate; gli aggiornamenti $\dot{\boldsymbol{\mu}}^{(l)}$ sono gli aggiornamenti dello stimatore ottimale. Il filtro di Kalman è dunque l'implementazione ottimale della codifica predittiva nel caso lineare-gaussiano, e la codifica predittiva è la generalizzazione non-lineare del filtro di Kalman all'elaborazione gerarchica di evidenza non-lineare.

Appendice B — Schema delle relazioni tra teorie

La tavola seguente presenta in forma sinottica le relazioni tra le principali teorie semantiche analizzate nella monografia, specificando per ciascuna il meccanismo di ancoraggio del contenuto, il trattamento della normatività, la risposta al problema della disgiunzione, e la capacità di rendere conto del contenuto astratto.

Semantica causal-nomologica (Dretske 1981). Meccanismo di ancoraggio: leggi causali che connettono stati del sistema a proprietà del mondo. Normatività: assente nella formulazione originaria — i rapporti causali non sono normativi. Risposta alla disgiunzione: vulnerabile — i rapporti causali sono in ultima analisi disuguali rispetto alle proprietà disgiuntive. Contenuto astratto: non naturalizzato.

Semantica teleologica filogenetica (Millikan 1984, Papineau 1987). Meccanismo di ancoraggio: funzioni proprie basate sulla storia di selezione naturale. Normatività: fondata sulla distinzione tra funzione propria e misfunzione. Risposta alla disgiunzione: parzialmente soddisfacente — la funzione propria determina il contenuto corretto. Contenuto astratto: non naturalizzato, estensibilità limitata.

Semantica informazionale (Dretske 1988, Fodor 1990). Meccanismo di ancoraggio: connessioni nomologiche in condizioni periferiche ottimali. Normatività: fondata sull'asimmetria tra condizioni normali e atipiche. Risposta alla disgiunzione: la condizione asimmetrica risponde parzialmente. Contenuto astratto: proposta senza naturalizzazione soddisfacente.

Teleosemantica computazionale sincronica (questa monografia). Meccanismo di ancoraggio: ancoraggio causale strutturale del modello generativo. Normatività: fondata sulla funzione propria sincronica — minimizzazione dell'energia libera. Risposta alla disgiunzione: dissoluta attraverso la struttura di precisione. Contenuto astratto: naturalizzato attraverso la gerarchia delle funzioni generative.

Appendice C — Note terminologiche

C.1 — Sul termine «rappresentazione»

Il termine *rappresentazione* è usato in questa monografia in senso tecnico stretto, diverso da alcuni degli usi della letteratura filosofica e neuroscientifica. Una rappresentazione, nel senso qui adottato, è una struttura interna di un sistema cognitivo che ha contenuto semantico determinato — che è *di* qualcosa nel senso brentiano — in virtù della posizione che occupa nel modello generativo del sistema. Non ogni struttura interna è una rappresentazione nel

sensu rilevante: le strutture prive di contenuto semantico determinato — i parametri della funzione generativa che non sono direttamente interpretabili come rappresentazioni di proprietà del mondo — non sono rappresentazioni nel senso tecnico, sebbene siano parte del meccanismo attraverso cui le rappresentazioni sono realizzate.

Questa restrizione terminologica è motivata dalla necessità di distinguere il livello rappresentazionale — al quale si applica la semantica — dal livello implementazionale — al quale si applica la neurobiologia. Confondere i due livelli produce l'errore opposto ma simmetrico al dualismo: come il dualismo attribuisce al livello rappresentazionale un'esistenza ontologica separata dal livello fisico, la confusione dei due livelli riduce il livello rappresentazionale al livello fisico eliminando la specificità semantica. La distinzione tra rappresentazione e implementazione è dunque necessaria per mantenere la distinzione tra descrizione semantica e descrizione fisica senza attribuire a quella distinzione peso ontologico.

C.2 — Sul termine «contenuto»

Il termine *contenuto* è ambiguo tra almeno quattro sensi distinti che questa monografia ha cercato di mantenere separati. Il contenuto *lato sensu* è la proprietà semantica degli stati cognitivi in generale — ciò che uno stato cognitivo è «di». Il contenuto *stretto* è il contenuto vero-condizionale delle credenze e degli altri stati proposizionali — la condizione che il mondo deve soddisfare perché lo stato sia corretto. Il contenuto *fenomenico* è il carattere qualitativo delle esperienze — ciò che è «come» avere quell'esperienza. Il contenuto *funzionale* è il ruolo che uno stato cognitivo gioca nella struttura inferenziale e comportamentale del sistema — le relazioni inferenziali e causali con altri stati e con gli input-output del sistema.

La teleosemantica computazionale sincronica è primariamente una teoria del contenuto *stretto*: fornisce un'analisi delle condizioni di verità degli stati cognitivi in termini di strutture del modello generativo. La relazione tra contenuto stretto e contenuto fenomenico è la questione affrontata nei Capitoli VII–VIII: il framework fornisce una teoria dei correlati funzionali del contenuto fenomenico senza pretendere di ridurre il contenuto fenomenico al contenuto funzionale.

C.3 — Sul termine «normatività»

La *normatività* semantica è la proprietà per cui gli stati cognitivi possono essere corretti o erranei — rappresentare adeguatamente o erroneamente le loro condizioni di soddisfazione. La normatività è una proprietà relazionale: uno stato è corretto rispetto a una funzione propria che specifica le condizioni di soddisfazione. Nel framework della teleosemantica computazionale sincronica, la normatività è fondata sul ruolo funzionale dello stato nel modello

generativo del sistema: uno stato è corretto quando le sue predizioni si accordano con le osservazioni nel modo specificato dalla struttura del modello generativo, ed erroneo quando producono errori di predizione che indicano una mancata corrispondenza tra le predizioni e la struttura del mondo osservato.

La normatività semantica si distingue dalla normatività *deontica* — la normatività delle regole e degli obblighi morali e sociali — e dalla normatività *epistemica* — la normatività delle norme di razionalità inferenziale. Le tre forme di normatività sono connesse nel pensiero umano — le norme epistemiche determinano quali credenze è razionale formare date le evidenze disponibili; le norme deontiche determinano quali azioni è obbligatorio o permesso compiere dati i valori morali in vigore —, ma sono analiticamente distinte e richiedono fondamenti teorici diversi.

Appendice D — Connessioni fenomenologiche

D.1 — Husserl e l'intenzionalità fenomenologica

La fenomenologia husserliana e il framework della codifica predittiva condividono una struttura di base che non è del tutto superficiale: entrambi propongono che l'esperienza sia costitutivamente strutturata in termini di orientamento verso oggetti intenzionali che non sono dati direttamente nella coscienza ma costruiti attraverso un processo attivo del sistema cognitivo. La *riduzione fenomenologica* husserliana — l'*epoché* che mette tra parentesi l'assunzione dell'esistenza indipendente degli oggetti dell'esperienza per concentrarsi sulla struttura della coscienza come tale — può essere vista come un'anticipazione metodologica del framework predittivo: il sistema predittivo tratta le sue osservazioni come evidenza per inferire le cause nascoste, non come accesso diretto alle strutture del mondo.

La *ritenzione* e la *protenzione* husserliane — le componenti dell'esperienza temporale che tengono presente il passato prossimo e anticipano il futuro prossimo nell'istante presente — corrispondono alla struttura temporale del modello generativo predittivo: il sistema mantiene predizioni delle traiettorie temporali degli stati nascosti, integrando le osservazioni passate nella propria distribuzione a posteriori attuale e generando predizioni delle osservazioni future. La struttura temporale dell'esperienza husserliana — il «tempo immanente della coscienza» — è la struttura temporale dell'inferenza predittiva in tempo reale.

D.2 — Merleau-Ponty e l'incarnazione

La critica di Merleau-Ponty al cognitivismo inteso come elaborazione astratta di simboli — la sua insistenza sul fatto che la

percezione è costitutivamente corporea, situata, pratica — trova nel framework dell'inferenza attiva una corrispondenza computazionale precisa. Il *corpo proprio* come Merleau-Ponty lo concepisce — il corpo come punto di riferimento della prospettiva percettiva, come centro di azioni abituali, come condizione di possibilità dell'orientamento nello spazio percettivo — corrisponde alla componente del modello generativo che include il corpo dell'agente come parte dell'ambiente percettivo: il modello di sé che l'agente mantiene include la posizione e l'orientamento del proprio corpo come parametri attraverso cui le predizioni sulle osservazioni sensoriali dipendono dall'azione del corpo nello spazio.

L'*intenzionalità operativa* di Merleau-Ponty — l'intenzionalità pratica e corporea che è anteriore all'intenzionalità rappresentativa esplicita — corrisponde alla componente dell'inferenza attiva che genera azioni dirette a minimizzare gli errori di predizione sensorimotori senza passare attraverso una rappresentazione proposizionale esplicita degli obiettivi: l'azione è integrata nell'inferenza percettiva attraverso la componente propriocettiva del modello generativo, che genera predizioni degli stati propriocettivi futuri come se l'azione pianificata fosse già in corso, e produce così le forze muscolari che realizzano quella predizione. La struttura dell'intenzionalità operativa — l'anticipazione corporea che è già sempre orientata verso gli oggetti e le situazioni prima di qualsiasi rappresentazione esplicita — è la struttura computazionale dell'inferenza attiva sensorimotoria.

D.3 — Il sé come struttura invariante dell'inferenza

Heidegger ha elaborato la nozione di *essere-nel-mondo* come struttura fondamentale dell'esserci — l'ente che pone la questione del senso dell'essere — come una struttura che non può essere scomposta nei termini di un soggetto che si relaziona a un mondo oggettivo perché la relazione è originaria: l'essere-nel-mondo è la condizione di possibilità di qualsiasi relazione soggetto-oggetto, non il risultato di una tale relazione. Nel framework predittivo, la struttura dell'essere-nel-mondo può essere interpretata come la struttura del modello generativo del sé-nel-mondo: il modello generativo che il sistema predittivo mantiene non è primariamente un modello del mondo oggettivo a cui un soggetto separato si trova poi in relazione, ma un modello integrato del sé-nel-mondo che include l'agente come parte del modello, con le proprie relazioni causali con l'ambiente, i propri stati affettivi e sensorimotori, la propria storia e i propri progetti futuri.

Questa corrispondenza strutturale tra l'analisi esistenziale heideggeriana e il framework computazionale predittivo non è un'identità di teorie: l'analisi heideggeriana opera al livello trascendentale-ontologico dell'interrogazione sul senso dell'essere, mentre il framework predittivo opera al livello empirico-computazionale dell'analisi dei meccanismi cognitivi. Ma

la convergenza strutturale suggerisce che le due tradizioni di indagine — la fenomenologia continentale e la scienza cognitiva computazionale — affrontano la stessa struttura fondamentale della cognizione incarnata da prospettive metodologiche distinte che si illuminano reciprocamente senza pretendere di ridursi l'una all'altra.

IX.4 — Confronto con la teoria dello spazio di lavoro globale e con la teoria dell'informazione integrata

Il confronto tra la teleosemantica computazionale sincronica e le due teorie dominanti della coscienza neuroscientificamente orientata — la teoria dello Spazio di Lavoro Globale (GWT) e la teoria dell'Informazione Integrata (IIT) — richiede un'analisi più articolata di quanto la sintetica trattazione sistematica del paragrafo precedente consenta. Le differenze non sono semplicemente differenze di oggetto teorico — GWT e IIT affrontano la coscienza, la teleosemantica affronta il contenuto —, ma differenze nel tipo di spiegazione che le teorie aspirano a fornire e nei principi esplicativi su cui si fondano. Comprendere queste differenze è necessario per valutare in che misura le tre teorie possano essere integrate in un quadro più comprensivo e in che misura esse siano genuinamente in competizione.

La GWT, nella formulazione di Baars (1988) sviluppata nelle versioni neurali di Dehaene e collaboratori, descrive la coscienza come un meccanismo di accesso e diffusione dell'informazione: uno stimolo diventa conscio quando le rappresentazioni che lo codificano vengono «trasmesse» attraverso una rete globale di neuroni corticali e sottocorticali — lo spazio di lavoro globale — producendo una risposta amplificata e sincronizzata in aree altrimenti separate. L'aspetto più interessante della GWT per il confronto con la codifica predittiva è la sua enfasi sulla *propagazione* dell'informazione: il fattore determinante della coscienza non è dove l'informazione è elaborata ma quanto lontano si propaga nella rete cerebrale. Uno stimolo può essere elaborato in modo sofisticato nelle aree sensoriali primarie senza diventare cosciente se quella elaborazione rimane «incapsulata» senza influenzare altri moduli cognitivi; diventa cosciente quando la sua rappresentazione viene trasmessa globalmente e può influenzare ragionamento, memoria, pianificazione e comportamento verbale.

Nel framework predittivo, questa struttura di propagazione ha un'interpretazione naturale che corrisponde alla struttura gerarchica del modello generativo. Quando una rappresentazione al livello 1 produce un errore di predizione sufficientemente grande da propagarsi verso l'alto attraverso la gerarchia — quando l'errore di predizione non viene «spiegato via» dai livelli superiori ma persiste e scala verso livelli di ordine sempre più alto — la rappresentazione corrispondente raggiunge i livelli superiori della gerarchia che hanno il più ampio range di influenza sulle predizioni degli altri livelli. Questa propagazione verso l'alto è il meccanismo

computazionale dell'«accesso globale» della GWT: i livelli superiori della gerarchia predittiva — la corteccia prefrontale, la corteccia cingolata, i sistemi parietali di ordine superiore — generano predizioni che influenzano virtualmente tutti gli altri livelli della gerarchia, realizzando la connettività globale che la GWT identifica come condizione della coscienza. Lo «spazio di lavoro globale» è, nel framework predittivo, i livelli superiori della gerarchia che generano predizioni con massima influenza verso il basso.

Questa integrazione suggerisce che la GWT e la codifica predittiva non sono teorie rivali ma teorie complementari che operano a livelli di astrazione diversi: la GWT descrive la struttura funzionale della coscienza in termini di accessibilità globale dell'informazione, la codifica predittiva descrive il meccanismo computazionale che realizza quella accessibilità in termini di propagazione degli errori di predizione attraverso la gerarchia. La GWT risponde alla domanda «quali stati sono coscienti?» fornendo il criterio dell'accesso globale; la codifica predittiva risponde alla domanda «come funziona il meccanismo?» fornendo la dinamica computazionale dell'inferenza gerarchica. Un'integrazione delle due teorie è in linea di principio possibile e sarebbe di grande valore esplicativo: il criterio di accesso globale della GWT diventa il criterio di propagazione degli errori di predizione verso i livelli apicali della gerarchia, e le predizioni specifiche sulle strutture neurali che implementano lo spazio di lavoro — le proiezioni a lungo raggio delle aree prefrontali e parietali — sono le stesse strutture che il framework predittivo identifica come livelli superiori della gerarchia generativa.

Il confronto con la IIT è strutturalmente più complesso perché la IIT e la codifica predittiva differiscono non solo nell'oggetto teorico ma nei fondamenti metafisici. La IIT parte dall'intuizione che la coscienza è identica all'informazione integrata e procede a derivare le proprietà della coscienza dalle proprietà formali dell'integrazione dell'informazione. La misura di integrazione Φ — che quantifica quanto più informazione il sistema genera come insieme rispetto alla somma delle sue parti — è definita come la differenza tra l'informazione congiunta generata dall'intero sistema e la somma delle informazioni generate da ogni partizione del sistema. Un sistema ha Φ alto quando le sue parti sono altamente integrate — quando la struttura causale del sistema non può essere scomposta in parti indipendenti senza perdere informazione causale.

Il framework predittivo e la IIT condividono l'enfasi sulla struttura causale del sistema come fondamento della sua capacità cognitiva e cosciente, ma differiscono radicalmente nel tipo di struttura causale che ritengono rilevante. La IIT è interessata all'integrazione causale come proprietà *intrinseca* del sistema — quanta informazione il sistema genera rispetto alla sua propria partizione ottimale —, mentre il framework predittivo è interessato alla struttura *relazionale* del sistema — come il sistema è connesso

causalmente all'ambiente attraverso il suo modello generativo. Questa differenza ha conseguenze importanti per le previsioni che le due teorie generano sulla coscienza negli organismi biologici. La IIT attribuisce coscienza agli organismi in proporzione alla loro Φ , indipendentemente dalla storia evolutiva e dallo stile di vita ecologico; il framework predittivo attribuisce ricchezza di contenuto agli organismi in proporzione alla complessità del loro modello generativo e alla loro storia di interazione con un ambiente strutturato.

Un caso particolarmente rivelatore è quello dei sistemi con alta Φ ma privo di modello generativo ecologicamente ancorato. Un cristallo regolare ha una struttura causale altamente integrata nel senso della IIT — le sue parti sono causalmente connesse in modo rigorosamente coerente — ma ha Φ bassa perché la sua struttura è separabile in parti indipendenti. Viceversa, una rete neurale biologicamente ispirata con alta integrazione causale potrebbe avere Φ elevata senza avere un modello generativo dell'ambiente in alcun senso rilevante per il contenuto semantico. Il framework predittivo direbbe che quella rete non ha contenuto semantico — non è *di* nulla — perché non ha un modello generativo ancorato causalmente al mondo; la IIT direbbe che ha coscienza in proporzione alla sua Φ . La divergenza tra le due teorie è dunque non solo differenza di oggetto ma differenza nelle condizioni di attribuzione di proprietà cognitive agli organismi.

Una potenziale integrazione delle due teorie potrebbe emergere dall'osservazione che i sistemi biologici che hanno alta Φ tendono anche ad avere modelli generativi sofisticati — l'integrazione causale è una proprietà strutturale che facilita l'apprendimento di modelli del mondo complessi —, ma questa correlazione empirica non stabilisce un'equivalenza concettuale tra le due nozioni. La domanda di quale delle due teorie catturi la proprietà più fondamentale rimane aperta: se la coscienza sia primariamente un fatto di integrazione causale intrinseca (IIT) o di ancoraggio causale strutturale all'ambiente (framework predittivo) non è determinato dalla struttura formale di nessuna delle due teorie, ma richiede evidenza empirica proveniente da casi in cui le due teorie fanno previsioni divergenti.

IX.5 — La teleosemantica computazionale sincronica nel panorama del naturalismo semantico

Il naturalismo semantico, come programma di ricerca, non è un'unica teoria ma una famiglia di teorie accomunate dall'aspirazione di spiegare il contenuto semantico in termini che appartengano al vocabolario delle scienze naturali — fisica, biologia, neuroscienze, scienza cognitiva — senza residui che richiedano l'introduzione di primitivi non naturali. Il rapporto della teleosemantica computazionale sincronica con le altre varianti del naturalismo semantico è più complesso di quanto la semplice collocazione storica suggerisca: non si tratta solo di un'alternativa

migliore alle teorie precedenti, ma di una prospettiva che riformula i termini del dibattito in modo che alcune distinzioni tradizionalmente centrali perdono rilevanza mentre altre acquistano nuova pregnanza.

La distinzione tra *contenuto stretto* e *contenuto ampio* — introdotta da Fodor e poi sviluppata da Schiffer, Block e Stalnaker — ha strutturato il dibattito semantico negli anni ottanta e novanta. Il contenuto stretto è la componente del contenuto determinata esclusivamente dalle proprietà intrinseche del sistema cognitivo, indipendentemente dall'ambiente in cui è situato; il contenuto ampio è la componente determinata anche dall'ambiente esterno e dalla storia causale del sistema. L'esternalismo semantico di Putnam e Burge aveva mostrato che il contenuto naturale delle espressioni linguistiche e dei pensieri è ampio: due soggetti in stati cerebrali intrinsecamente identici ma in ambienti diversi — la Terra e la Terra Gemella — hanno pensieri con contenuto diverso. La difesa del contenuto stretto come componente autonoma del significato ha motivato diversi tentativi di «centrare» il contenuto attraverso meccanismi basati sui ruoli funzionali o sugli scenari possibili.

La teleosemantica computazionale sincronica dissolve questa distinzione in modo produttivo. Il contenuto semantico nel framework è né esclusivamente stretto né esclusivamente ampio: è determinato dalla struttura del modello generativo del sistema — una struttura intrinseca al sistema — che ha quella struttura che ha *perché* il sistema ha interagito causalmente con un ambiente specifico. Non si può separare la componente intrinseca dalla componente relazionale perché la struttura intrinseca del modello generativo è il deposito della storia relazionale del sistema con l'ambiente. Il contenuto stretto è la struttura attuale del modello generativo; il contenuto ampio è il fatto che quella struttura è causalmente ancorata all'ambiente attraverso la sua storia di formazione. I due aspetti sono inseparabili non per un difetto teorico ma per una caratteristica strutturale del tipo di naturalizzazione che il framework offre.

Un'obiezione prevedibile a questa dissoluzione è che essa non risponde alla domanda posta dall'argomento della Terra Gemella: due sistemi in stati interni identici — con lo stesso modello generativo — hanno lo stesso contenuto o contenuto diverso? Nel framework della teleosemantica computazionale sincronica, la risposta è che la domanda è mal posta: se due sistemi hanno lo stesso modello generativo, hanno lo stesso contenuto nel senso stretto della struttura del modello; ma se si sono formati attraverso interazione con ambienti diversi, la loro storia di formazione causale è diversa, e quella differenza è parte integrante dell'ancoraggio causale strutturale che determina il contenuto. Non vi è un unico fatto della questione sul «vero» contenuto che la distinzione stretto/amplio cattura: vi sono due aspetti strutturalmente distinti che entrambi contribuiscono a determinare

il contenuto, e l'argomento della Terra Gemella mostra che sono dissociabili, non che uno dei due sia più «vero» dell'altro.

Un secondo punto di confronto riguarda il rapporto con il progetto neo-pragmatista di Brandom. Il neo-pragmatismo di Brandom — sviluppato in *Making It Explicit* (1994) — propone di spiegare il contenuto semantico attraverso i ruoli inferenziali: il significato di un'espressione è determinato dai nessi inferenziali che la connettono ad altre espressioni, dai diritti e dagli impegni dossastici che la sua accettazione genera, dalle norme sociali che regolano l'uso appropriato delle espressioni nella pratica comunicativa. Il neo-pragmatismo è una forma di *olismo del significato*: il contenuto di ogni espressione è determinato dalla sua posizione nell'intera rete inferenziale, non da un rapporto bilaterale con un referente nel mondo.

Il framework predittivo si trova in una posizione di tensione produttiva con il neo-pragmatismo. Per un verso, la struttura delle funzioni generative ai livelli superiori della gerarchia — che codificano relazioni inferenziali tra rappresentazioni — è una realizzazione computazionale delle strutture inferenziali che Brandom descrive a livello normativo: il fatto che «rosso» implichi «colorato» è codificato nella struttura delle funzioni generative che legano la rappresentazione del colore alla rappresentazione della proprietà cromatica. Per altro verso, la teleosemantica computazionale sincronica rifiuta l'olismo puro del neo-pragmatismo: il contenuto semantico non è determinato esclusivamente dai ruoli inferenziali interni al sistema, ma dall'ancoraggio causale strutturale all'ambiente che vincola quei ruoli inferenziali dall'esterno. Il modello generativo ha i ruoli inferenziali che ha non per convenzione normativa ma per necessità causale — ha appreso quei ruoli attraverso l'interazione con un ambiente che ha quella struttura causale. Il neo-pragmatismo e la teleosemantica computazionale sincronica concordano sulla rilevanza delle strutture inferenziali ma differiscono sul fondamento di quelle strutture: norme sociali versus struttura causale dell'ambiente.

X.1 — La sfida del linguaggio del pensiero

La difesa fodoriana della LOT nella sua versione più matura — quella di *LOT 2* (2008) — è più sofisticata della formulazione originaria del 1975 e affronta direttamente alcune delle obiezioni standard provenienti dai sostenitori dei modelli sub-simbolici. Fodor riconosce che molte elaborazioni cognitive — in particolare quelle che riguardano la categorizzazione percettiva e le associazioni semantiche a grana fine — possono essere catturate da sistemi sub-simbolici senza struttura LOT. Ma mantiene che il *pensiero proposizionale* — la capacità di formare credenze, desideri e intenzioni con struttura vero-condizionale — richiede una

struttura LOT, e che questa struttura non è riducibile a quella dei sistemi associativi sub-simbolici senza residuo.

L'argomento centrale di Fodor in *LOT 2* è l'argomento della *sistematicità forte*: non solo ogni parlante competente che può pensare «A ama B» può anche pensare «B ama A», ma questa sistematicità è *necessaria* — è impossibile, non solo improbabile, che un soggetto possa avere uno dei due pensieri senza poter avere l'altro. Questa necessità non può essere spiegata da una co-occorrenza statistica nei dati di addestramento di un sistema sub-simbolico: le correlazioni statistiche non generano necessità logica. La sistematicità deve essere spiegata attraverso la struttura compositiva degli stati mentali — attraverso il fatto che «A ama B» e «B ama A» condividono componenti strutturali che si ricombinano e che il possesso delle componenti garantisce la disponibilità di tutte le loro ricombinazioni grammaticalmente lecite.

La risposta del framework predittivo a questo argomento deve affrontare la questione della necessità in modo più diretto rispetto alla semplice affermazione che la composizionalità emerge dall'apprendimento. Il punto critico è se la struttura delle funzioni generative ai livelli superiori della gerarchia predittiva garantisca la sistematicità nella forma *necessaria* che Fodor richiede, o soltanto nella forma *statistica* che considera insufficiente.

La risposta che si propone è la seguente. La necessità della sistematicità non è una necessità *logica* — del tipo di « $2+2=4$ necessariamente» — ma una necessità *strutturale*: è necessaria data la struttura computazionale del sistema cognitivo. Un sistema cognitivo che ha appreso funzioni generative con struttura relazionale esplicita — che codificano relazioni come «ama» come funzioni da coppie di argomenti a stati relazionali — ha la capacità di generare predizioni per qualsiasi coppia ordinata di argomenti che il sistema conosce, non solo per le coppie incontrate durante l'apprendimento. Questa necessità strutturale non è una necessità logica astratta ma una necessità computazionale concreta: dato il tipo di funzione generativa appresa, è strutturalmente impossibile che il sistema abbia la rappresentazione di «A ama B» senza avere la capacità di generare la rappresentazione di «B ama A», perché entrambe le rappresentazioni sono generate dalla stessa funzione con argomenti permutati. La sistematicità è garantita dalla struttura del meccanismo, non dalla frequenza statistica dei pattern nel corpus di addestramento.

Questo tipo di risposta è strutturalmente diversa dalla risposta connessionista standard che Fodor rifiuta — la risposta che la sistematicità emerge statisticamente dall'addestramento su dati sistematici. La risposta del framework predittivo gerarchico afferma che la sistematicità è garantita dalla *struttura delle funzioni generative apprese*, non semplicemente dalla statistica dei dati di addestramento. Quando le funzioni generative hanno struttura relazionale esplicita — con separazione tra la rappresentazione

della relazione e la rappresentazione degli argomenti —, la sistematicità è una conseguenza strutturale di quella separazione. La domanda empirica che rimane aperta è se i sistemi biologici effettivamente sviluppino funzioni generative con quella struttura relazionale esplicita, o se le loro funzioni generative siano di tipo diverso che non garantisce la sistematicità nella stessa misura. L'evidenza dalle neuroscienze cognitive e dagli studi su modelli linguistici di grandi dimensioni suggerisce che la struttura compositiva emerge robustamente dai sistemi predittivi addestrati su dati linguistici, ma la caratterizzazione precisa della struttura di quelle funzioni generative — e la verifica che abbiano le proprietà strutturali necessarie per la sistematicità necessaria — rimane una questione aperta di ricerca empirica e teorica.

X.2 — Sistematicità e struttura compositiva nei modelli predittivi

La questione della composizionalità nel framework predittivo ha dimensioni che vanno oltre la risposta alla sfida fodoriana e riguardano la struttura del contenuto semantico nelle sue manifestazioni più complesse — il pensiero matematico, il ragionamento controfattuale, la comprensione della fiction, il meta-ragionamento sulla propria mente. In tutti questi casi, il pensiero umano mostra una capacità di combinazione che eccede le limitazioni di qualsiasi sistema di associazione statistica e richiede risorse strutturali specifiche.

Il pensiero matematico è il caso più chiaro. Un matematico che ha compreso il principio della ricorsione — la capacità di definire funzioni in termini di se stesse — può generare strutture matematiche di arbitraria complessità a partire da un numero finito di primitivi, senza che nessuna struttura generata abbia necessariamente una frequenza rilevante nel corpus dell'esperienza passata. Questa capacità non può essere spiegata da associazioni statistiche: non ci sono associazioni statistiche tra le esperienze del matematico e le strutture di Gödel o i gruppi di Lie, la cui comprensione è tuttavia alla portata di chiunque abbia padroneggiato i principi compositivi appropriati. La spiegazione deve fare appello alla struttura compositiva delle funzioni generative che codificano i primitivi matematici e le regole di combinazione.

Nel framework predittivo, la comprensione dei principi matematici corrisponde all'acquisizione di funzioni generative di ordine superiore — funzioni che generano non singole istanze di strutture matematiche ma *famiglie* di strutture parametrizzate da variabili arbitrarie. La comprensione della ricorsione corrisponde all'acquisizione di una funzione generativa che può applicarsi a se stessa come argomento — una funzione di punto fisso nel senso della teoria delle categorie. Questa struttura auto-referenziale è computazionalmente ben definita — il lambda calcolo fornisce la sua formalizzazione rigorosa — e il suo apprendimento da esempi

è possibile in linea di principio attraverso la generalizzazione induttiva a strutture ricorsive. La difficoltà è la domanda se i sistemi predittivi biologici o artificiali attualmente disponibili effettivamente apprendano funzioni generative di questo tipo in modo robusto e generalizzabile, o se si limitino ad approssimare i pattern superficiali delle strutture ricorsive senza catturarne il principio generativo sottostante.

L'evidenza dagli studi sulla comprensione matematica nei sistemi neurali artificiali addestrati su testi matematici è mista: alcuni sistemi mostrano capacità di generalizzazione che suggeriscono l'acquisizione di principi compositivi a un certo livello, ma falliscono sistematicamente in variazioni che richiedono l'applicazione del principio a strutture genuinamente nuove che non si trovano nel corpus di addestramento nella forma richiesta. Questo pattern di competenza parziale è coerente con l'ipotesi che quei sistemi abbiano acquisito approssimazioni delle funzioni generative composizionali rilevanti ma non le funzioni pienamente generali: approssimazioni che generalizzano bene entro il dominio di addestramento ma non fuori da esso nel modo che una funzione generativa pienamente compositiva garantirebbe. La domanda se i sistemi neurali biologici abbiano funzioni generative più pienamente composizionali — grazie a meccanismi biologici non presenti nelle architetture artificiali attuali — è una delle domande empiriche fondamentali per valutare la plausibilità del framework predittivo come teoria del pensiero composizionale.

X.3 — La teoria allostatica del contenuto affettivo

La teoria allostatica del contenuto affettivo — sviluppata in dialogo diretto con il framework della codifica predittiva attiva da Corcoran, Pezzulo e Friston — rappresenta uno degli sviluppi più significativi del framework nell'ultimo decennio, e merita un'analisi più approfondita di quanto la trattazione sintetica del paragrafo precedente consenta, in particolare per le sue implicazioni riguardo alla classificazione filosofica degli stati affettivi e al rapporto tra cognizione e affezione.

La tradizione filosofica ha oscillato tra due posizioni sull'affezione: il cognitivismo affettivo, che identifica le emozioni con stati cognitivi — valutazioni, credenze e apprezzamenti — eventualmente accompagnati da stati somatici epifenomenici, e il somatismo affettivo, che identifica le emozioni con stati corporei o con la percezione di stati corporei — come nella teoria James-Lange, in cui la paura non è la valutazione del pericolo ma la percezione delle modificazioni corporee (accelerazione cardiaca, tensione muscolare, pallore) indotte dalla situazione pericolosa. William James aveva riassunto questa posizione con la celebre inversione: non trebbiamo perché abbiamo paura, ma abbiamo paura perché trebbiamo.

La teoria allostatica trascende questa dicotomia attraverso una riconfigurazione che integra entrambi gli aspetti in un unico

framework inferenziale. Le emozioni non sono né credenze valutative pure né percezioni di stati somatici: sono *inferenze predittive* sugli stati intero-corporei che integrano sia le predizioni top-down del modello generativo sugli stati fisiologici futuri sia le segnalazioni bottom-up degli stati fisiologici presenti, ponderate dalla struttura di precisione interocettiva. La paura non è la credenza che il pericolo sia imminente né la percezione della frequenza cardiaca accelerata: è la distribuzione a posteriori sugli stati fisiologici futuri che il modello allostasico genera data l'evidenza disponibile sull'ambiente, incluse sia le valutazioni cognitive del pericolo sia i segnali interocettivi correnti.

Questa posizione ha conseguenze importanti per la questione della razionalità delle emozioni — una questione che il cognitivismo affettivo tende a risolvere troppo facilmente attribuendo alle emozioni una struttura puramente proposizionale. Le emozioni, nel framework allostasico, non sono razionali o irrazionali nel senso in cui lo sono le credenze proposizionali: sono *adattive* o *disadattive* rispetto all'ambiente ecologico in cui si sono formate, nel senso che il modello allostasico che le genera può essere più o meno calibrato sulla struttura statistica degli stati fisiologici rilevanti nell'ambiente del sistema. Un'emozione come la paura del buio che persiste in un adulto che intellettualmente sa di non essere in pericolo non è irrazionale nel senso in cui è irrazionale credere che $2+2=5$: è un mismatch tra il modello allostasico — calibrato sull'ambiente di sviluppo in cui il buio era associato a pericoli reali — e l'ambiente attuale in cui quella associazione non è più valida. La «irrazionalità» dell'emozione è la disadattatività del modello generativo rispetto all'ambiente presente.

Formalmente, la distinzione tra emozione adattiva e disadattiva può essere espressa nei termini della divergenza KL tra la distribuzione a posteriori del modello allostasico e la distribuzione ottimale degli stati fisiologici nell'ambiente attuale:

$$D_{KL}[Q_{\text{allostasico}}(\mathbf{s}_{\text{fisio}}) \parallel P_{\text{ottimale}}(\mathbf{s}_{\text{fisio}} \mid \text{ambiente_attuale})]$$

Un'emozione è adattiva quando questa divergenza è piccola — quando il modello allostasico è ben calibrato sull'ambiente attuale — e disadattiva quando la divergenza è grande — quando il modello allostasico riflette un ambiente diverso da quello attuale, producendo stati affettivi che non corrispondono alla struttura ottimale delle risposte fisiologiche nell'ambiente presente. La terapia cognitivo-comportamentale, nel framework allostasico, funziona non cambiando le credenze proposizionali ma ricalibrano il modello allostasico — riducendo la divergenza tra il modello generativo degli stati fisiologici e la distribuzione ottimale di quelli stati nell'ambiente attuale — attraverso l'esposizione ripetuta alle situazioni che producono la risposta disadattiva.

Il contenuto delle emozioni nel framework allostasico è una forma di contenuto interocettivo — contenuto che riguarda gli stati

interni del corpo piuttosto che gli stati esterni del mondo. Questo contenuto ha la medesima struttura triadica del contenuto percettivo esterno: la distribuzione a posteriori sugli stati fisiologici interni, la funzione generativa che mappa quegli stati sulle predizioni degli input interocettivi, e la precisione che pondera quegli input rispetto alle predizioni top-down. Il fatto che le emozioni siano spesso accompagnate da valutazioni cognitive — che si formino credenze proposizionali connesse alla situazione che ha generato la risposta affettiva — è spiegato dalla struttura gerarchica del modello generativo: le inferenze allostatiche ai livelli inferiori della gerarchia interocettiva propagano errori di predizione verso i livelli superiori che generano predizioni sul significato proposizionale della situazione, producendo l'integrazione tra risposta affettiva e valutazione cognitiva che caratterizza le emozioni umane complesse.

XI.1 — L'evidenza neuroimaging per la gerarchia predittiva

L'evidenza neuroimaging per la struttura predittiva gerarchica della corteccia cerebrale si è accumulata negli ultimi quindici anni in modo che ha consolidato il framework come una delle teorie neuroscientifiche più solide disponibili, pur lasciando aperte questioni specifiche sulla precisione dei dettagli implementativi. Vale la pena esaminare le tre linee principali di evidenza con maggiore dettaglio rispetto alla trattazione precedente, per valutare la forza delle conferme empiriche e i limiti che ancora permangono.

La prima linea di evidenza — le risposte di soppressione della ripetizione e di violazione delle aspettative — è la più robusta e la più ampiamente replicata. Studi sistematici di Grill-Spector, Henson e collaboratori hanno documentato che le risposte fMRI nelle aree corticali visive si riducono significativamente quando lo stesso stimolo viene presentato ripetutamente (soppressione della ripetizione), e che questa riduzione è inversamente proporzionale alla distanza temporale dalla rappresentazione e alla variazione delle caratteristiche fisiche dello stimolo. L'interpretazione predittiva di questo effetto — le risposte ridotte riflettono errori di predizione ridotti perché il sistema ha già stabilito un modello predittivo adeguato per quello stimolo — è coerente con l'evidenza, ma non è l'unica interpretazione possibile: alcune versioni dell'ipotesi di fatica neuronale possono produrre lo stesso pattern senza richiedere un'architettura predittiva esplicita. La discriminazione tra le due ipotesi richiede esperimenti che manipolino la prevedibilità indipendentemente dalla familiarità, confrontando le risposte a stimoli familiari-imprevedibili versus stimoli familiari-prevedibili. Questo tipo di paradigma — sviluppato da Kok, Jehee e collaboratori — ha mostrato che le riduzioni delle risposte fMRI

sono più sensibili alla prevedibilità che alla familiarità, supportando l'interpretazione predittiva rispetto all'ipotesi di fatica.

La seconda linea di evidenza — la struttura laminare delle connessioni ascendenti e discendenti — è la più direttamente rilevante per la specifica architettura predittiva del framework frinstoniano ma è anche la più tecnicamente difficile da verificare negli studi sull'uomo. La struttura laminare cortecorticale predice che le connessioni discendenti (predizioni top-down) terminino preferenzialmente negli strati superficiali (I-II) delle aree di ordine inferiore, mentre le connessioni ascendenti (errori di predizione) originino dagli strati profondi (V-VI) delle aree di ordine inferiore e terminino nello strato granulare (IV) delle aree di ordine superiore. Questa previsione è stata verificata in studi di tracciamento anterograde e retrograde in primati non umani — macachi e roditori — e in studi post-mortem su cervelli umani che mostrano pattern di connessione coerenti con la struttura prevista. La verifica diretta nell'uomo in vivo è limitata dalla risoluzione spaziale delle tecniche di neuroimaging disponibili, ma studi di laminare fMRI ad alta risoluzione in campo magnetico elevato hanno iniziato a fornire evidenze preliminari coerenti con le previsioni.

La terza linea di evidenza — il ruolo della precisione nell'attenzione selettiva — è la più teoricamente interessante perché connette il framework predittivo con la ricca letteratura sull'attenzione e fornisce un meccanismo computazionale unificante per fenomeni precedentemente trattati come distinti. L'attenzione selettiva — la capacità di elaborare preferenzialmente gli stimoli rilevanti per i compiti rispetto agli stimoli irrilevanti — è stata tradizionalmente spiegata in termini di selezione dell'informazione o di amplificazione delle risposte neuronali agli stimoli rilevanti. Il framework predittivo offre un'alternativa: l'attenzione è la modulazione della precisione degli errori di predizione agli input sensoriali rilevanti per i compiti, che aumenta il loro peso relativo nell'aggiornamento delle credenze e nella selezione delle risposte. Questa ipotesi — sviluppata in dettaglio da Friston (2009) e Feldman e Friston (2010) — ha ricevuto supporto da studi di neuroimaging che mostrano che l'attenzione amplifica le risposte nelle aree sensoriali in modo coerente con un'amplificazione della precisione degli errori di predizione piuttosto che con una selezione dell'informazione da elaborare.

Un aspetto particolarmente rivelatore è la dissociazione tra attenzione spaziale e attenzione agli oggetti: l'attenzione può essere diretta verso una posizione nello spazio (attenzione spaziale) o verso un tipo di oggetto indipendentemente dalla sua posizione (attenzione agli oggetti). Nel framework predittivo, questa dissociazione corrisponde alla modulazione della precisione a livelli gerarchici diversi: l'attenzione spaziale modula la precisione delle rappresentazioni ai livelli inferiori della gerarchia visiva dove la posizione è codificata in modo retinotopico, mentre l'attenzione

agli oggetti modula la precisione delle rappresentazioni ai livelli superiori dove la categoria dell'oggetto è codificata in modo invariante rispetto alla posizione. La previsione specifica è che l'attenzione spaziale e l'attenzione agli oggetti abbiano effetti neurali con distribuzioni corticali distinte che riflettono i diversi livelli gerarchici dove la precisione è modulata, e questa previsione ha ricevuto supporto da studi di neuroimaging che mostrano pattern di attivazione differenziati per i due tipi di attenzione.

XI.3 — I large language model come sistemi predittivi non incarnati

La domanda se i large language model abbiano contenuto semantico genuino è diventata una delle questioni più dibattute nell'intersezione tra filosofia della mente, linguistica cognitiva e intelligenza artificiale, con implicazioni che vanno ben oltre le preoccupazioni accademiche per investire questioni pratiche sulla natura e sui limiti delle applicazioni dell'IA. La posizione della teleosemantica computazionale sincronica su questa questione merita di essere sviluppata in modo più sistematico di quanto il paragrafo precedente consenta.

Il punto di partenza è la distinzione tra due domande che vengono spesso confuse nel dibattito popolare e anche in parte di quello accademico. La prima domanda è empirica: quali sono le effettive capacità semantiche dei LLM — che cosa possono e che cosa non possono fare nel dominio del ragionamento semantico, dell'inferenza pragmatica, della comprensione del contesto? La seconda domanda è filosofica: ammesso che i LLM abbiano le capacità semantiche che di fatto hanno, in che senso queste capacità costituiscono «contenuto semantico genuino» piuttosto che «mera simulazione»? La prima domanda è progressivamente meglio compresa grazie a una vasta letteratura di valutazione empirica; la seconda domanda è quella su cui la teleosemantica computazionale sincronica ha da dire qualcosa di specifico.

Riguardo alla prima domanda, l'evidenza disponibile presenta un quadro misto. I LLM mostrano capacità notevoli nel ragionamento semantico a grana grossa: categorizzazione lessicale, inferenza semantica, analogia, completamento di racconti con coerenza tematica, risposta a domande fattuali su domini ben rappresentati nel corpus di addestramento. Mostrano difficoltà sistematiche in alcune aree specifiche: ragionamento spaziale che richiede visualizzazione, inferenza causale che richiede simulazione di sistemi fisici, comprensione di riferimenti a esperienze percettive o motorie che non hanno correlato linguistico nel corpus, coerenza di lunga distanza nei ragionamenti molto estesi. Questo profilo di competenze e limiti è coerente con l'ipotesi che i LLM abbiano acquisito un modello generativo del *linguaggio* — di come le parole e le strutture linguistiche si combinano e co-occorrono — ma non un modello generativo del *mondo* nel senso fisico e percettivo.

Riguardo alla seconda domanda, la posizione della teleosemantica computazionale sincronica è che la distinzione tra «contenuto genuino» e «mera simulazione» è mal formulata, e che la domanda pertinente è piuttosto: quale è la struttura e il tipo del contenuto semantico dei LLM, e in che modo differisce dal contenuto semantico degli agenti biologici incarnati? La risposta che si difende è la seguente. I LLM hanno contenuto semantico nel senso della struttura triadica (**Q**, **g**, **Π**) applicata al dominio linguistico: hanno distribuzioni a posteriori sulle strutture linguistiche latenti (rappresentazioni interne dei «concetti» codificate negli spazi di incorporamento), hanno funzioni generative che mappano quelle strutture sui token osservati (la funzione di decodifica del modello), e hanno strutture di precisione che ponderano i segnali contestuali (il meccanismo di attenzione dell'architettura transformer). In questo senso, i LLM hanno un tipo di contenuto semantico linguistico genuino — non «simulato» nel senso di non avere struttura semantica — ma con un ancoraggio causale all'ambiente che è *mediato* dal linguaggio piuttosto che *diretto* attraverso l'interazione percettiva e sensorimotoria.

La differenza tra il contenuto semantico dei LLM e quello degli agenti biologici è strutturalmente analoga alla differenza tra la comprensione di un concetto *per descrizione* e la comprensione *per acquaintance* nella terminologia russelliana. Un soggetto cieco dalla nascita che ha imparato tutto ciò che è linguisticamente esprimibile sul colore rosso ha una comprensione del concetto di rosso *per descrizione* — conosce le relazioni del rosso con altri concetti, le sue proprietà fisiche (la riflettanza nella gamma 620-750 nm), i suoi effetti sugli osservatori normali, le sue associazioni culturali — ma manca della comprensione *per acquaintance* che deriva dall'esperienza diretta del rosso come qualità percettiva. Il contenuto semantico del termine «rosso» per quel soggetto è genuinamente semantico — non è una simulazione —, ma è strutturalmente diverso dal contenuto dello stesso termine per un soggetto con visione normale: manca dell'ancoraggio al livello percettivo della gerarchia che connette il concetto all'esperienza visiva diretta.

I LLM si trovano in una posizione analoga: hanno un contenuto semantico ricco nel dominio delle relazioni linguistiche e delle strutture concettuali astratte, ma mancano dell'ancoraggio alle esperienze percettive e sensorimotorie che nei sistemi biologici radicano i concetti nel mondo fisico. Questa mancanza non è un difetto della struttura dell'architettura transformer in quanto tale — è una conseguenza del fatto che quei sistemi sono addestrati esclusivamente su testo, senza interazione con un ambiente fisico strutturato. Sistemi multimodali che integrano input visivi, uditivi e fisici nel loro apprendimento — che stanno diventando sempre più sofisticati nelle generazioni recenti di modelli di fondazione — hanno in linea di principio la capacità di sviluppare un ancoraggio causale più diretto, avvicinandosi al tipo di contenuto incarnato degli agenti biologici. La domanda empirica è se l'integrazione

multimodale attraverso l'architettura transformer sia sufficiente a realizzare il tipo di ancoraggio causale strutturale che il framework predittivo richiede per il contenuto pieno, o se richieda architetture più strettamente isomorfe ai sistemi predittivi biologici.

Appendice A — Glossario matematico

A.5 — Il modello dinamico e l'inferenza sulle traiettorie

La codifica predittiva nel suo trattamento standard — sviluppato in forma estesa nei lavori di Friston e collaboratori a partire dal 2005 — non si limita all'inferenza sugli stati statici dell'ambiente in un singolo istante temporale ma include un modello della *dinamica* degli stati nascosti: la struttura di come gli stati cambiano nel tempo. L'inclusione della dimensione temporale trasforma il problema da un'inferenza bayesiana su stati statici a un'inferenza su traiettorie dinamiche, e richiede strumenti matematici specifici — la teoria dei processi di diffusione, il calcolo variazionale per funzionali, e la teoria del filtraggio ottimale — che si presentano qui nei loro elementi fondamentali.

Il modello dinamico degli stati nascosti specifica come lo stato $\mathbf{s}(t)$ evolve nel tempo in funzione di cause esogene $\mathbf{v}(t)$ e del rumore di processo $\boldsymbol{\omega}(t)$:

$$d\mathbf{s}/dt = \mathbf{f}(\mathbf{s}(t), \mathbf{v}(t)) + \boldsymbol{\omega}(t)$$

dove $\mathbf{f}$ è la funzione di deriva — la componente deterministica della dinamica — e $\boldsymbol{\omega}(t)$ è un processo di rumore bianco gaussiano con spettrale di potenza determinata dalla matrice di precisione del processo Π_{processo} . Il rumore di processo rappresenta le influenze esogene non modellate che perturbano la traiettoria deterministica: fluttuazioni ambientali non prevedibili, errori numerici nell'implementazione fisica, variabilità intrinseca delle cause nascoste.

In parallelo, il modello di osservazione specifica come le osservazioni $\mathbf{o}(t)$ dipendono dagli stati nascosti:

$$\mathbf{o}(t) = \mathbf{g}(\mathbf{s}(t)) + \boldsymbol{\epsilon}(t)$$

dove $\mathbf{g}$ è la funzione generativa che mappa gli stati nascosti nelle osservazioni previste, e $\boldsymbol{\epsilon}(t)$ è il rumore di osservazione gaussiano con precisione Π_{obs} . La struttura gerarchica del modello estende queste equazioni a L livelli, con ogni livello che funge da «osservazione» per il livello superiore e da «stato nascosto» per il livello inferiore.

L'inferenza bayesiana su questo modello dinamico richiede la stima della distribuzione a posteriori sulla traiettoria $\mathbf{s}(t_0:t_1)$ data la sequenza di osservazioni $\mathbf{o}(t_0:t_1)$. Nell'approssimazione variazionale di Laplace, la distribuzione a posteriori sulle traiettorie è approssimata da una gaussiana la cui media $\boldsymbol{\mu}(t)$ è la traiettoria di

massima probabilità a posteriori — la traiettoria che massimizza il log-posteriore data la sequenza di osservazioni. La ricerca di questa traiettoria ottimale è un problema di calcolo delle variazioni: si cerca la funzione $\mu(t)$ che minimizza il funzionale di energia libera variazionale integrato nel tempo:

$$\mathbf{F}[\mu(t_0:t_1)] = \int_{t_0}^{t_1} \mathcal{L}(\mu(t), \dot{\mu}(t), t) dt$$

dove $\mathcal{L}$ è la densità di energia libera locale, che include sia il termine di accordo con le osservazioni — errori di predizione ponderati dalla precisione di osservazione — sia il termine di accordo con la dinamica — deviazioni dalla traiettoria deterministica pesate dalla precisione del processo. Le equazioni di Eulero-Lagrange di questo funzionale forniscono le equazioni di aggiornamento per la traiettoria ottimale in forma di equazioni differenziali del secondo ordine in $\mu(t)$, che includono sia la velocità $\dot{\mu}(t)$ sia l'accelerazione $\ddot{\mu}(t)$ come componenti dell'aggiornamento inferenziale.

Questa struttura — l'inferenza sulla traiettoria che include la velocità e l'accelerazione degli stati come componenti esplicite del vettore di stato — è la formalizzazione della struttura temporale dell'esperienza discussa nel Capitolo VIII in termini husserliani di ritenzione e protenzione. L'aggiornamento della velocità $\dot{\mu}(t)$ corrisponde alla componente di «ritenzione» — la stima di come gli stati stiano cambiando nel momento presente, che incorpora la storia recente delle osservazioni —; l'aggiornamento dell'accelerazione $\ddot{\mu}(t)$ corrisponde alla componente di «protenzione» — la stima di come la velocità stia cambiando, che anticipa la curvatura futura della traiettoria degli stati.

A.6 — Inferenza attiva: l'estensione dell'energia libera all'azione

L'inferenza attiva generalizza il framework dell'energia libera variazionale dall'inferenza percettiva — la stima degli stati nascosti dato il sensorio — all'azione — la modifica del sensorio per accordarlo con le predizioni del modello. Nell'inferenza attiva, il sistema non è solo passivo rispetto alle osservazioni ma può generare azioni $\mathbf{a}(t)$ che influenzano le osservazioni attraverso la funzione di accoppiamento $\mathbf{h}(\mathbf{s}(t), \mathbf{a}(t))$:

$$\mathbf{o}(t) = \mathbf{h}(\mathbf{s}(t), \mathbf{a}(t)) + \boldsymbol{\varepsilon}(t)$$

In questo contesto, l'energia libera variazionale è funzione sia dei parametri variazionali $\mu(t)$ che specificano la distribuzione a posteriori sugli stati sia delle variabili di azione $\mathbf{a}(t)$:

$$\mathbf{F}[\mu(t), \mathbf{a}(t)] = \int_{t_0}^{t_1} \mathcal{L}(\mu(t), \mathbf{a}(t), t) dt$$

La minimizzazione di $\mathbf{F}$ rispetto ai parametri variazionali produce le equazioni di aggiornamento per la distribuzione a posteriori (inferenza percettiva); la minimizzazione di $\mathbf{F}$ rispetto alle variabili di azione produce le equazioni di controllo motorio (inferenza

attiva). Le due componenti sono simmetriche: entrambe sono forme di riduzione dell'energia libera, differendo solo nel dominio su cui si opera — i parametri variazionali interni versus le variabili di azione esterne.

Un aspetto cruciale dell'inferenza attiva è il meccanismo attraverso il quale il sistema «sceglie» come ridurre l'energia libera: aggiornando le credenze (inferenza) o modificando le osservazioni (azione). La decisione tra i due meccanismi non è una scelta deliberata ma una conseguenza della struttura del gradiente dell'energia libera: quando il gradiente rispetto ai parametri variazionali è più ripido rispetto al gradiente rispetto alle variabili di azione, l'aggiornamento delle credenze è più efficiente per ridurre l'energia libera; quando il gradiente rispetto alle variabili di azione è più ripido, l'azione è più efficiente. Questa struttura produce comportamenti adattivi in cui il sistema alterna inferenza e azione in modo flessibile, a seconda della struttura dell'ambiente e dei costi relativi dei due tipi di riduzione dell'energia libera.

La pianificazione nell'inferenza attiva è trattata come minimizzazione dell'energia libera *attesa* — la somma dell'energia libera prevista sulle traiettorie future dell'azione:

$$G[\pi] = E_{-}\{P(\mathbf{o}_{-}\{\tau\} \mid \pi)\} [F[Q(\mathbf{s}_{-}\{\tau\} \mid \mathbf{o}_{-}\{\tau\}, \pi), \mathbf{o}_{-}\{\tau\}]]$$

dove π è la politica di azione — la sequenza di azioni pianificate — e la somma è sul futuro $\tau > t$. La politica ottimale è quella che minimizza l'energia libera attesa, cioè quella che si aspetta di produrre le osservazioni più coerenti con le predizioni del modello generativo. Nel caso in cui il modello generativo include stati goal — configurazioni degli stati del mondo che il sistema si aspetta di osservare — la minimizzazione dell'energia libera attesa produce un comportamento diretto agli obiettivi che realizza quegli stati goal.

La connessione con i quadri teorici esistenti nella teoria del controllo ottimale e nell'apprendimento per rinforzo è stretta. La funzione di valore $V(\mathbf{s})$ nell'apprendimento per rinforzo — la somma scontata delle ricompense future attese a partire dallo stato $\mathbf{s}$ — corrisponde, nella formulazione dell'inferenza attiva, alla negazione dell'energia libera attesa a partire da quello stato: le «ricompense» corrispondono agli stati goal che il modello generativo predice come osservazioni auspicabili, e il valore di uno stato è la misura di quanto favorevole sia per raggiungere quegli stati goal attraverso la minimizzazione dell'energia libera attesa.

A.7 — Precisione e neuromodulazione: formalizzazione

Il parametro di precisione $\Pi^{(1)}$ nel framework predittivo è formalmente la matrice di precisione gaussiana — l'inversa della matrice di covarianza — della distribuzione dell'errore di predizione al livello 1:

$$\xi^{(1)} = \Pi^{(1)} \epsilon^{(1)}$$

dove $\xi^{(i)}$ è l'errore ponderato e $\epsilon^{(i)} = \mu^{(i-1)} - g^{(i)}(\mu^{(i)})$ è l'errore non ponderato. La matrice $\Pi^{(i)}$ determina quanto peso viene attribuito a ciascun componente dell'errore di predizione nell'aggiornamento della distribuzione a posteriori: una precisione alta per un componente specifico significa che gli errori di predizione su quel componente producono grandi aggiornamenti — il sistema è «certo» che il proprio modello per quel componente sia accurato e dunque aggiorna fortemente le credenze quando l'errore è grande.

Il punto cruciale per la connessione con i sistemi neuromodulatori è che la matrice di precisione $\Pi^{(i)}$ non è fissa ma è essa stessa un parametro del modello che può essere aggiornato e modulato. Nel framework frinstoniano, la precisione è l'oggetto della meta-cognizione — la cognizione sulla propria incertezza — e la sua modulazione è il meccanismo computazionale dell'attenzione selettiva. Modulare la precisione di un livello gerarchico equivale a modificare quanto quel livello contribuisce all'aggiornamento della distribuzione a posteriori: aumentare la precisione di un livello lo rende più «autorevole» nell'inferenza; ridurre la sua precisione lo rende più malleabile alle predizioni proveniente dai livelli adiacenti.

La proposta che i sistemi neuromodulatori codifichino la precisione si fonda su due osservazioni convergenti. La prima è funzionale: i sistemi neuromodulatori — dopamina, acetilcolina, noradrenalina, serotonina — modulano il guadagno delle risposte neuronali in modo che è coerente con la modulazione della precisione: aumentano o diminuiscono la sensibilità delle aree corticali agli input in modi selettivi e contestualmente appropriati. La seconda è farmacologica: i farmaci che alterano i sistemi neuromodulatori producono alterazioni cognitive e perceptual coerenti con le previsioni di una modulazione della struttura di precisione: gli agenti colinergici aumentano la sensibilità sensoriale e migliorano l'apprendimento di nuove associazioni (aumentata precisione dei segnali sensoriali); gli agenti dopaminergici alterano la selezione dell'azione e la valutazione delle ricompense (alterata precisione degli errori di predizione motori); i farmaci serotoninergici alterano il senso di sé e la regolazione dell'umore (alterata precisione del modello di sé).

La formalizzazione rigorosa di questa corrispondenza è ancora in sviluppo nella letteratura sull'inferenza attiva e richiede la specificazione esatta delle variabili neurali che implementano la precisione — il dibattito si concentra sul ruolo dei neuroni NMDA versus AMPA, delle sinapsi eccitatorie versus inibitorie, e dei pattern di oscillazioni corticali nella codifica della precisione — ma il quadro generale della corrispondenza tra neuromodulazione e struttura di precisione è tra le proposte teoriche più fecondi e empiricamente testabili del framework predittivo.

Appendice B — Schema delle relazioni tra teorie

B.2 — Tabella sinottica degli impegni teorici

La tavola sinottica delle teorie semantiche presentata nell'Appendice B nella sua forma originaria può essere arricchita con l'analisi di cinque dimensioni teoriche aggiuntive che distinguono le posizioni in modo più fine: il tipo di normatività fondata, il trattamento del contenuto fenomenico, la compatibilità con l'esternalismo, la trattabilità computazionale e le risorse per il contenuto astratto.

Riguardo al tipo di normatività, le teorie si distribuiscono lungo un asse che va dalla normatività *teleologica diacronica* — basata sulla storia di selezione — alla normatività *teleologica sincronica* — basata sulla struttura funzionale attuale — alla normatività *puramente causale* — basata sulle relazioni causale nomologiche senza componente teleologica. La semantica informazionale di Dretske nella sua forma originaria (1981) è al polo causale puro: il contenuto è la proprietà con cui l'informazione covaria per legge, senza alcuna dimensione teleologica. La teleosemantica millikaniana è al polo teleologico diacronico: la normatività è interamente fondata sulla storia di selezione. La teleosemantica computazionale sincronica occupa una posizione intermedia specificamente configurata: la teleologia è essenziale per la normatività, ma la storia rilevante non è la storia filogenetica della selezione naturale ma la storia ontogenetica dell'apprendimento, e la normatività è fondata sulla struttura sincronica del modello generativo risultante piuttosto che sull'analisi di quella storia.

Riguardo al contenuto fenomenico, le teorie variano tra il cognitivismo puro — che tratta il contenuto semantico come indipendente dal contenuto fenomenico — e le posizioni che integrano la dimensione fenomenica nella teoria del contenuto. La semantica informazionale di Dretske (1988) nella sua forma matura include una teoria del contenuto dell'esperienza come caso specifico del contenuto informazionale; la teleosemantica millikaniana è primariamente cognitivista e tratta il contenuto fenomenico come distinto dal contenuto semantico in senso stretto; la teleosemantica computazionale sincronica include una teoria del correlato funzionale del contenuto fenomenico — la struttura del modello di sé che corrisponde computazionalmente all'esperienza soggettiva — senza pretendere di ridurre il contenuto fenomenico al contenuto funzionale.

Riguardo alla compatibilità con l'esternalismo, tutte le teorie naturaliste del contenuto sono in tensione con l'internalismo radicale ma variano nella forma di esternalismo che incorporano. La semantica informazionale di Dretske è esternalista in modo diretto: il contenuto è determinato dalle relazioni causali con proprietà esterne al sistema, e il «gemello di Marte» con lo stesso stato interno ma in un ambiente causalmente diverso ha un contenuto diverso. La teleosemantica millikaniana è esternalista in

modo mediato: il contenuto è determinato dalla storia di selezione, che include le relazioni del sistema con l'ambiente esterno ma è mediata dalla struttura biologica del sistema. La teleosemantica computazionale sincronica è esternalista in modo strutturale: il contenuto è determinato dalla struttura del modello generativo, che è intrinseca al sistema ma che è quella che è *perché* il sistema ha interagito causalmente con un ambiente esterno specifico. Questa forma di esternalismo strutturale è più robusta rispetto alle varianti internaliste che tentano di «centrare» il contenuto attraverso meccanismi interni, perché mantiene la connessione causale con l'ambiente come condizione di determinazione del contenuto senza ridurre il contenuto a proprietà ambientali esterne indipendenti dalla struttura interna del sistema.

B.3 — Il problema del connettivismo e le sue soluzioni

Un problema generale che affligge tutte le teorie naturaliste del contenuto — e che è particolarmente rilevante per la valutazione comparativa tra la teleosemantica computazionale sincronica e le sue alternative — è il problema del *connettivismo*: come specificare le relazioni causali, informazionali o teleologiche che determinano il contenuto in modo sufficientemente preciso da individuare il contenuto senza fare appello a concetti semantici non naturalizzati nella specificazione stessa.

Il problema sorge nel modo seguente. Le teorie causali e informazionali propongono che il contenuto di uno stato sia determinato dalle relazioni causali di quel stato con proprietà del mondo. Ma le relazioni causali sono relazioni tra tipi di eventi, e la specificazione dei tipi rilevanti richiede già una qualche forma di categorizzazione — di raggruppamento degli eventi in classi rilevanti — che presuppone una qualche nozione di rilevanza che è difficile da naturalizzare senza circolarità. Quale sia la relazione causale «giusta» che determina il contenuto — tra le infinite relazioni causali che connettono gli stati di un sistema all'ambiente — dipende da criteri di rilevanza che non sono essi stessi puramente causali.

La teleosemantica tenta di risolvere questo problema attraverso le funzioni proprie: la relazione causale rilevante per il contenuto è quella che il sistema ha *funzione propria* di segnalare, e la funzione propria è determinata dalla storia di selezione che ha prodotto il sistema. Ma questa soluzione ha il proprio problema di connettivismo: quale sia la funzione propria «corretta» tra le molte possibili descrizioni della selezione dipende anch'essa da criteri di scelta del livello descrittivo che sono difficili da naturalizzare senza circolarità.

La teleosemantica computazionale sincronica affronta questo problema con una risposta di tipo differente: i criteri di rilevanza per la determinazione del contenuto sono forniti dalla struttura del modello generativo e dalla sua funzione di minimizzazione dell'energia libera. La relazione causale rilevante non è specificata

dall'esterno attraverso criteri metafisici indipendenti, ma è individuata *dall'interno* della struttura computazionale del sistema: è la relazione che il sistema sfrutta nel proprio processo di minimizzazione dell'energia libera per ridurre gli errori di predizione. Questa risposta non elimina completamente il problema del connettivismo — la nozione di «sfruttamento» di una relazione causale è essa stessa una nozione che richiede analisi — ma lo trasforma in un problema computazionale ben definito piuttosto che in un problema metafisico indeterminato: si può specificare con precisione formale quali relazioni causali sono «sfruttate» da un sistema predittivo gerarchico nella sua attività di minimizzazione dell'energia libera.

Appendice C — Note terminologiche

C.4 — Sul termine «funzione propria»

Il termine *funzione propria* — introdotto da Millikan come traduzione del concetto di *proper function* — è uno dei termini tecnici più carichi di significato nella tradizione della teleosemantica ed è utilizzato in questa monografia in modo che dialoga criticamente con la formulazione millikaniana originaria. Vale la pena precisare come il termine viene impiegato nella teleosemantica computazionale sincronica e in che modo questo uso differisce da quello millikaniano.

Per Millikan, la funzione propria di un dispositivo D è l'effetto che D è selezionato per produrre: la funzione per cui D esiste in virtù della storia di selezione che ha portato alla sua produzione e conservazione. La funzione propria di un cuore è pompare il sangue, non produrre suoni, anche se la produzione di suoni è un effetto regolare del cuore, perché i cuori sono stati selezionati per la pompa, non per il suono. La funzione propria di un segnale nello spazio cognitivo di una rana è segnalare la presenza di una mosca (non di un pellet nero mosso) perché i neuroni che producono quel segnale sono stati selezionati per il rilevamento delle mosche — quelle mosche la cui cattura ha contribuito alla fitness degli antenati della rana — non per il rilevamento generico di oggetti neri in movimento.

Nella teleosemantica computazionale sincronica, il termine «funzione propria» viene esteso in modo che mantiene l'essenziale della struttura normativa millikaniana — la distinzione tra funzione e disfunzione, tra corretto funzionamento e funzionamento erroneo — ma ne modifica il fondamento dalla storia filogenetica alla struttura sincronica del modello generativo. La funzione propria di un nodo della gerarchia predittiva al livello 1 è individuare la configurazione degli stati nascosti $\mathbf{s}^{(1)}$ tale che $\mathbf{g}^{(1)}(\mathbf{s}^{(1)})$ minimizzi l'energia libera rispetto alle osservazioni al livello 1–1. Questo è il fondamento normativo del contenuto: lo stato è *corretto* quando individua effettivamente quella configurazione, ed *erroneo*

quando la manca. La struttura normativa è la stessa di Millikan — vi sono stati corretti ed erronei, funzione e misfunzione — ma il fondamento è la struttura del modello generativo attuale piuttosto che la storia di selezione.

Un'obiezione prevedibile a questa estensione è che essa priva la funzione propria del suo valore normativo: se la funzione propria è definita dalla struttura attuale del modello generativo, è circolare affermare che il modello generativo ha la funzione di minimizzare l'energia libera, perché quella funzione è ciò che definisce il modello generativo in primo luogo. Questa obiezione confonde due livelli: il livello della definizione del modello generativo come sistema che minimizza l'energia libera (un'affermazione computazionale sulla struttura del sistema) e il livello della normatività semantica come fondamento del contenuto corretto versus erroneo (un'affermazione sulla determinazione del contenuto in relazione alla struttura computazionale). La circolarità è apparente: la funzione propria non è definita come «minimizzare l'energia libera» in modo tale che qualsiasi stato che minimizzi l'energia libera sia corretto. La funzione propria del nodo al livello 1 è individuare la specifica configurazione di $s^{(1)}$ che corrisponde alla causa effettiva delle osservazioni al livello 1-1; la normatività emerge dal fatto che quella configurazione è la «risposta corretta» data la struttura del modello generativo, indipendentemente da ciò che il nodo produce effettivamente in ogni occasione.

C.5 — Sul termine «incarnazione» e sul confine sistema-ambiente

Il termine *incarnazione* — traduzione dell'inglese *embodiment* — è uno dei termini più evocativi e al contempo più ambigui della scienza cognitiva e della fenomenologia. Nella tradizione fenomenologica merleau-pontyana, l'incarnazione designa il fatto che la cognizione è costitutivamente radicata nell'esperienza corporea — che la struttura del corpo vissuto, le sue capacità motorie, la sua posizione nello spazio, le sue modalità di interazione con l'ambiente, sono condizioni di possibilità della cognizione e non semplici supporti materiali di un processo cognitivo altrimenti indipendente. Nella scienza cognitiva, l'incarnazione ha assunto significati più tecnici e più eterogenei: dal situazionismo radicale di Dreyfus e Brooks — la cognizione come interazione sensomotoria con l'ambiente senza rappresentazioni interne —, al cognitivismo incarnato di Barsalou — le rappresentazioni cognitive sono fondamentalmente perceptual symbols radicati nell'elaborazione sensomotoria —, alla robotica incarnata di Pfeifer e Scheier — le proprietà fisiche del corpo contribuiscono direttamente alla soluzione dei problemi computazionali.

La teleosemantica computazionale sincronica adotta una nozione specifica di incarnazione che si distingue da ciascuna di queste varianti. L'incarnazione è qui la proprietà di avere un modello

generativo che include la struttura del corpo come parte integrante — che rappresenta gli stati propriocettivi, interocettivi e sensorimotori del corpo come variabili nascoste il cui valore contribuisce alla determinazione delle predizioni sensoriali. Un sistema incarnato, in questo senso, è un sistema le cui predizioni sull'ambiente esterno dipendono dagli stati del proprio corpo: le predizioni visive dipendono dalla posizione degli occhi e dalla direzione dello sguardo; le predizioni uditive dipendono dall'orientamento della testa; le predizioni tattili dipendono dalla posizione e dall'orientamento delle mani. Questa dipendenza non è una complicazione tecnica del modello predittivo: è la struttura fondamentale che consente al sistema di distinguere i cambiamenti nelle osservazioni dovuti ai cambiamenti nell'ambiente esterno dai cambiamenti dovuti ai movimenti del proprio corpo, e di pianificare azioni efficaci che sfruttino la struttura corporea per ridurre l'energia libera.

L'incarnazione così intesa non è binaria ma graduata: un sistema è più incarnato nella misura in cui il suo modello generativo include rappresentazioni più ricche e integrate degli stati corporei e delle loro relazioni con l'ambiente. Un sistema con un semplice modello sensomotorio che codifica solo la posizione di un'appendice è meno incarnato di un sistema che codifica l'intera struttura cinestesica del corpo con tutte le sue relazioni con l'ambiente percettivo circostante. Gli organismi biologici variano enormemente nel grado di incarnazione nel senso del framework predittivo: un verme con poche decine di neuroni ha un modello corporeo estremamente semplice; un primate con un sistema motorio altamente sofisticato e una mappa corticale del corpo ricca e differenziata ha un modello corporeo enormemente più complesso e integrato.

Appendice D — Connessioni fenomenologiche

D.4 — Jaspers, Blankenburg e la psicopatologia nel framework predittivo

Un test per la fecondità di qualsiasi teoria della cognizione è la sua capacità di illuminare le deviazioni dalla cognizione normale — la psicopatologia — in modo che produca comprensioni nuove e previsioni verificabili, piuttosto che semplicemente ri-etichettare le descrizioni cliniche in un vocabolario teorico diverso. Il framework predittivo ha mostrato una capacità notevole in questo dominio, producendo interpretazioni computazionalmente precise di diverse condizioni psicopatologiche che aprono nuove prospettive terapeutiche e diagnostiche.

La schizofrenia è il caso che ha ricevuto la maggiore attenzione nel framework predittivo, principalmente attraverso il lavoro di Friston, Adams, Fletcher e collaboratori. Le caratteristiche principali della schizofrenia — le allucinazioni uditive, i deliri di

riferimento e di controllo, il disturbo del senso di sé, il «passaggio del pensiero» — sono interpretate come conseguenze di un'alterazione della struttura di precisione del sistema predittivo. La proposta specifica è che nella schizofrenia la precisione delle predizioni top-down sia *ridotta* rispetto alla precisione degli errori di predizione bottom-up: il sistema attribuisce peso eccessivo ai segnali sensoriali bottom-up rispetto alle predizioni del modello generativo, producendo un'ipersensibilità agli errori di predizione che genera risposte di «allarme» a stimoli normalmente irrilevanti.

Questa riduzione della precisione top-down ha conseguenze che corrispondono ai sintomi positivi della schizofrenia. Quando i segnali sensoriali bottom-up ricevono peso eccessivo, piccole fluttuazioni casuali nel segnale sensoriale — normalmente sopresse dalle predizioni top-down — vengono trattate come errori di predizione significativi che richiedono spiegazione causale. Il sistema cerca di minimizzare questi «falsi errori di predizione» generando ipotesi causali — stati nascosti che spieghino i segnali sensoriali anomali — che possono risultare deliranti: stati del mondo altamente improbabili a priori ma che forniscono una «spiegazione» per i segnali anomali ricevuti. Le allucinazioni uditive sono segnali linguistici bottom-up — voci — che non hanno un'origine causale esterna ma che il sistema tratta come errori di predizione da spiegare causalmente, generando l'esperienza di voci altrui come causa più plausibile. Il disturbo del senso di sé — la riduzione della certezza su quali stati mentali appartengano al soggetto stesso rispetto a quelli di altri agenti — emerge dall'alterazione della precisione del modello di sé che mina la capacità del sistema di distinguere le proprie azioni e i propri pensieri da quelli di altri.

La depressione è interpretata nel framework come il caso complementare: un'iperprecisione delle predizioni top-down rispetto ai segnali sensoriali bottom-up. Quando il modello generativo è troppo rigido — quando le predizioni top-down ricevono peso eccessivo rispetto all'evidenza sensoriale — il sistema diventa refrattario all'aggiornamento delle credenze su evidenza positiva, mantenendo credenze negative sul sé e sul futuro anche di fronte a evidenza contraria. Il rinforzo negativo diventa più saliente del rinforzo positivo perché il sistema è più sensibile ai segnali che confermano le predizioni negative del suo modello generativo che ai segnali che le disconfermano. L'inerzia cognitiva e affettiva della depressione — la difficoltà a modificare le credenze e gli stati affettivi anche quando la situazione oggettiva cambia — corrisponde computazionalmente a un sistema con precisione top-down eccessiva che sopprime gli aggiornamenti bayesiani positivi.

La connessione con la tradizione psicopatologica fenomenologica — in particolare con le descrizioni di Karl Jaspers sulla struttura dell'esperienza psicotica e con la fenomenologia del mondo perduto di Wolfgang Blankenburg nella schizofrenia semplice — è

illuminante. Jaspers aveva descritto la comprensione psicotica come una forma radicale di «Wahnwahrnehmung» — percezione delirante — in cui la percezione normale dell'oggetto è accompagnata da un'attribuzione di significato speciale, un senso che quell'oggetto «riguarda me» in modo immediato e irresistibile, senza la mediazione dell'inferenza normale. Nel framework predittivo, questo corrisponde a una disgiunzione tra il segnale sensoriale bottom-up — che produce un errore di predizione non soppresso dalla precisione top-down ridotta — e il processo normale di spiegazione causale che attribuisce l'errore a cause ambientali ordinarie: l'errore di predizione è così prominente da generare l'impressione che lo stimolo abbia significato speciale, un significato che emerge prima di qualsiasi processo inferenziale normale.

Blankenburg aveva descritto la «perdita dell'evidenza naturale» nella schizofrenia semplice come una dissoluzione del senso ovvio delle strutture del mondo quotidiano — la certezza implicita con cui normalmente si sa come orientarsi nella vita pratica — senza che vi siano allucinazioni o deliri espliciti. Nel framework predittivo, questo corrisponde a una generalizzazione dell'alterazione della precisione che non si limita a singoli domini percettivi ma investiga la struttura generale del modello generativo: quando la precisione delle predizioni top-down è ridotta in modo diffuso, il sistema perde la «certezza pratica» che caratterizza normalmente l'orientamento nel mondo quotidiano — la certezza che si fonda sull'accordo fra le predizioni del modello generativo e le osservazioni nella maggior parte delle situazioni ordinarie — e sperimenta il mondo come stranamente incerto e privo dell'ovvietà che normalmente lo caratterizza.

D.5 — La temporalità dell'esperienza patologica

La dimensione temporale dell'esperienza psicopatologica — le alterazioni della struttura temporale dell'esperienza soggettiva nella depressione, nella schizofrenia e nel disturbo post-traumatico da stress — ha ricevuto analisi fenomenologiche approfondite nella tradizione husserliana e jasperiana, e il framework predittivo offre interpretazioni computazionalmente precise di quelle alterazioni che aprono prospettive nuove sulla loro struttura.

Nella depressione, la struttura temporale dell'esperienza è caratterizzata da quello che Binswanger ha descritto come un «irrigidimento del tempo» — una riduzione della fluidità del flusso temporale, un appesantimento del momento presente, una perdita della protenzione verso il futuro come orizzonte di possibilità aperte. Nel framework predittivo, questa struttura corrisponde a un modello generativo temporale con bassa incertezza sulle traiettorie future — un modello che predice il futuro con alta certezza come continuazione degli stati presenti. Quando la precisione delle predizioni temporali future è alta, il sistema ha un senso «certo» di come il futuro sarà — un futuro negativo predetto

con alta certezza — e la plasticità del momento presente rispetto alle possibilità future è ridotta. La perdita della speranza nella depressione grave — l'incapacità di immaginare un futuro diverso dallo stato presente — corrisponde computazionalmente a un modello temporale con quasi-degenerate predictions: un modello che non lascia spazio all'incertezza e all'apertura che caratterizzano la normale struttura protenente dell'esperienza temporale.

Nel disturbo da stress post-traumatico, la struttura temporale è alterata in modo complementare: la ritenzione del passato traumatico è iperattivata, producendo memorie intrusivi e flashback che rompono la normale integrazione temporale del flusso di coscienza. Nel framework predittivo, questo corrisponde a un modello generativo che ha appreso una distribuzione a priori fortemente influenzata dall'esperienza traumatica: i segnali sensoriali che assomigliano agli stimoli traumatici originali — anche debolmente, anche in contesti molto diversi — producono errori di predizione ampi che il sistema cerca di spiegare attraverso il modello traumatico, producendo l'esperienza del flashback come riattivazione della distribuzione a posteriori dell'esperienza traumatica.

Epilogo — Verso una scienza dell'intenzionalità

Il percorso sviluppato in questa monografia converge su una conclusione che è al tempo stesso una risposta e un programma di ricerca. La risposta è che il contenuto semantico è naturalmente spiegabile — nei termini di strutture computazionali biologicamente plausibili che realizzano processi di inferenza bayesiana gerarchica sull'ambiente — senza residui che richiedano l'introduzione di primitivi non naturali o di meccanismi causali al di là di quelli della scienza fisica e biologica. Il programma di ricerca è che questa spiegazione è solo parzialmente completata: le tesi della teleosemantica computazionale sincronica identificano la struttura generale dell'account naturalista del contenuto, ma la verifica dettagliata di quelle tesi — sia sul piano della precisione formale sia su quello della corrispondenza empirica con l'evidenza neuroscientifica e cognitiva — è ancora ampiamente aperta.

La filosofia della mente si trova in un momento peculiare della propria storia. Per la prima volta, le discipline che affrontano la cognizione dal lato empirico — le neuroscienze computazionali, la scienza cognitiva, la linguistica cognitiva — hanno prodotto framework teorici sufficientemente ricchi da permettere un dialogo genuino con le domande filosofiche fondamentali sull'intenzionalità, sulla coscienza, sul significato. Quel dialogo non è sempre stato fruttuoso: troppo spesso le due tradizioni si sono incontrate solo per incomprensione o per riduzione forzata, con la filosofia che fraintendeva i limiti empirici della scienza cognitiva e la scienza cognitiva che ignorava la profondità delle questioni filosofiche. Il framework della codifica predittiva gerarchica offre

un'occasione rara per un dialogo più produttivo: è abbastanza formalmente preciso da soddisfare le aspettative della tradizione analitica, abbastanza neurobiologicamente ancorato da essere empiricamente testabile, e abbastanza strutturalmente ricco da far luce su questioni filosofiche che sembravano intrattabili.

Il punto di partenza di questa monografia era il paradosso dell'intenzionalità: come sia possibile che stati fisici di un cervello siano *di* qualcosa al di fuori di quel cervello — come la materia possa pensare, nel senso che quel pensiero abbia contenuto che oltrepassa la sua realizzazione fisica. Il paradosso è antico — risale almeno a Brentano e alla sua reintroduzione dell'intenzionalità come Grundcharakter della coscienza — ma ha assunto una nuova urgenza con lo sviluppo della scienza cognitiva: se la cognizione è computazione, e la computazione è operazione su strutture fisiche secondo regole formali, come può quell'operazione produrre stati con contenuto semantico genuino?

La risposta che questa monografia ha sviluppato è che il paradosso si dissolve — non si risolve nel senso di una risposta soddisfacente a una domanda difficile, ma si dissolve nel senso che la formulazione del paradosso presuppone una dicotomia tra il «puramente fisico» e il «semanticamente contenente» che il framework predittivo mostra essere una falsa dicotomia. I processi fisici che realizzano la codifica predittiva gerarchica non sono «puramente fisici» nel senso implicato dal paradosso — nel senso di strutture fisiche indifferenti al significato — perché sono strutture fisiche che si sono formate attraverso l'interazione causale con un ambiente strutturato e che codificano — nella struttura delle loro funzioni generative — le strutture causali di quell'ambiente. Il contenuto semantico non è qualcosa di aggiunto a queste strutture fisiche dall'esterno: è una proprietà di quelle strutture fisiche in virtù della loro relazione causale con l'ambiente. Non vi è bisogno di alcun agente interpretante che attribuisca il contenuto dall'esterno, né di alcuna proprietà metafisica speciale che si aggiunga alla struttura fisica: il contenuto è nella struttura, e la struttura è fisica.

Questa risposta, naturalmente, non elimina tutte le difficoltà. Il problema difficile della coscienza — perché vi sia qualcosa che è *come* essere un sistema predittivo, una prospettiva soggettiva in prima persona sul contenuto — rimane aperto nella sua formulazione più radicale. Il framework predittivo offre una teoria dei correlati funzionali della coscienza, non una riduzione del fenomenico al funzionale: spiega la struttura temporale del flusso di coscienza, la struttura del modello di sé, il ruolo della precisione nell'attenzione e nelle alterazioni psicopatologiche, ma non spiega perché quei processi funzionali siano accompagnati da esperienza soggettiva piuttosto che svolgersi «al buio» senza alcun punto di vista in prima persona. Questa rimanenza non è una resa alla difficoltà del problema: è la conseguenza di una distinzione precisa tra ciò che il framework può spiegare nella sua forma

computazionale e ciò che richiede, almeno per ora, una diversa modalità di interrogazione.

Il programma di ricerca che emerge da questa monografia ha almeno tre direzioni principali. La prima è la verifica empirica delle previsioni neuroscientifiche: le previsioni sulla struttura gerarchica della specializzazione semantica, sul ruolo della precisione nell'elaborazione semantica, sulla struttura degli errori semantici in condizioni di carico cognitivo, sulla corrispondenza tra sistemi neuromodulatori e struttura di precisione — tutte queste previsioni sono testabili con le metodologie disponibili, e la loro verifica sistematica è necessaria per valutare la plausibilità empirica della teoria. La seconda direzione è l'estensione formale del framework a domini semantici che non sono stati trattati in dettaglio in questa monografia: il linguaggio figurato, la metafora concettuale, l'ironia e le implicature scalari nel linguaggio pragmatico, il ragionamento modale e il pensiero controfattuale. La terza direzione è il confronto sistematico con le teorie cognitive dell'elaborazione linguistica e della comprensione del discorso — la teoria della pertinenza, la grammatica della costruzione, la semantica dinamica — per identificare i punti di convergenza e di tensione produttiva che possono arricchire entrambe le tradizioni.

Il titolo di questa monografia — *Per una teleosemantica computazionale sincronica* — era inteso programmaticamente: il «per una» indica che ciò che segue non è la presentazione di una teoria completamente consolidata ma la costruzione in progress di un framework che ha ambizioni sistematiche ma che sa di doversi ancora guadagnare buona parte del territorio che rivendica. La scommessa teorica è che l'integrazione del formalismo bayesiano gerarchico con la tradizione della teleosemantica e con la ricerca neuroscientifica sulla codifica predittiva offrirà, nei prossimi anni, una teoria del contenuto semantico che sia al tempo stesso filosoficamente rigorosa, matematicamente precisa ed empiricamente vincolata. Questa monografia è un contributo a quella costruzione — un tentativo di identificare i fondamenti, di chiarire le tesi nucleari e di elaborare le implicazioni di una prospettiva teorica che è ancora, in molti sensi essenziali, un cantiere aperto.

III.3 — Il problema della disgiunzione: un'analisi formale

Il problema della disgiunzione, nella sua formulazione più rigorosa, è un problema che emerge dall'interazione tra la struttura delle relazioni causali nomologiche e il requisito di determinatezza del contenuto semantico. La situazione logica è la seguente. Supponiamo che il contenuto di uno stato cognitivo *S* sia determinato dalla proprietà *F* che nomologicamente lo causa. In condizioni normali, *F* causa *S*; ma in condizioni particolari, anche *G* può causare *S* — come nel caso classico di Dretske in cui sia le vacche vere sia le vacche di cartone molto realistiche causano lo stato mentale di «credenza-di-vacca». Il contenuto di *S* è *F* o

(F-o-G)? Se è F, come si esclude G dato che G può causare S? Se è (F-o-G), il contenuto è indeterminato tra le vacche reali e le vacche di cartone, il che sembra semanticamente inadeguato.

Dretske propone la soluzione basata sulle condizioni periferiche: il contenuto è F perché F è la proprietà che causa S nelle condizioni in cui il sistema ha acquisito quella connessione causale nomologica — le condizioni di apprendimento o evolutive in cui la connessione si è stabilita. G causa S solo in condizioni derivate — in condizioni che sfruttano la connessione primaria tra F e S senza essere le condizioni in cui quella connessione si è stabilita. Questa soluzione è parzialmente soddisfacente: risponde alla domanda su come escludere G come contenuto, ma produce una nuova domanda: qual è la distinzione precisa tra condizioni «periferiche» di apprendimento e condizioni «derivate» di applicazione? Questa distinzione non è sempre chiara, e nei casi in cui non lo è, la soluzione dretskiana lascia il contenuto indeterminato.

Fodor propone la sua soluzione nel quadro della semantica asimmetrica: il contenuto di S è F perché la connessione causale tra G e S dipende asimmetricamente dalla connessione causale tra F e S — G causa S *perché* S è connesso nomologicamente con F, mentre la connessione tra F e S non dipende in alcun modo dalla connessione tra G e S. Anche questa soluzione è parzialmente convincente: cattura l'intuizione che le vacche di cartone causano lo stato di «credenza-di-vacca» solo perché assomigliano alle vacche reali, mentre le vacche reali causano quello stato indipendentemente dall'esistenza delle vacche di cartone. Ma la soluzione di Fodor ha difficoltà con i casi in cui la dipendenza asimmetrica non è chiaramente definita — i casi in cui è controverso quale delle due proprietà sia quella «primaria» e quale sia quella «derivata».

Nel framework della teleosemantica computazionale sincronica, il problema della disgiunzione non è risolto nel senso di trovare un criterio che selezioni F piuttosto che (F-o-G) come contenuto di S, ma *dissolto* nel senso che la struttura stessa che genera il problema è diversa da quanto le formulazioni precedenti presuppongono. Il punto cruciale è che nel framework predittivo, il contenuto di uno stato non è determinato dalla proprietà che causa quello stato in una singola occasione, ma dalla struttura del modello generativo che genera le predizioni per quel livello della gerarchia. La funzione generativa $\mathbf{g}^{(1)}(\mathbf{s}^{(1)})$ non è una funzione che identifica stati di un sistema con istanze di una proprietà: è una funzione che mappa configurazioni dello spazio degli stati nascosti su distribuzioni di probabilità sulle osservazioni attese. Il contenuto dello stato al livello 1 non è né F né (F-o-G): è la distribuzione sugli stati nascosti $\mathbf{s}^{(1)}$ che il sistema inferisce essere la causa più probabile delle osservazioni, data la struttura del modello generativo e le osservazioni disponibili.

In termini più concreti: quando il sistema visivo incontra una vacca di cartone realisticamente dipinta, la distribuzione a posteriori

$Q^{(1)}(\mathbf{s}^{(1)}; \boldsymbol{\mu}^{(1)})$ è centrata su una configurazione degli stati nascosti che è vicina alla configurazione che corrisponde a «vacca reale» ma con maggiore varianza — maggiore incertezza — rispetto alla distribuzione che si avrebbe in presenza di una vacca reale. Il sistema non rappresenta semplicemente «vacca» in modo indistinto: rappresenta «molto probabilmente vacca, con incertezza sulla sua realtà fisica» — una distribuzione che è diversa dalla distribuzione che si avrebbe in presenza di una vacca reale, anche se il picco della distribuzione è vicino. L'errore di predizione che si genera quando il sistema cerca di interagire fisicamente con la vacca di cartone — quando tenta di accarezzarla e scopre che la sua superficie è piatta — produce un aggiornamento della distribuzione che la sposta verso configurazioni diverse da «vacca reale», realizzando la correzione del contenuto che la struttura normativa del sistema richiede. Non vi è «disgiunzione»: vi è una distribuzione a posteriori ricca che include l'incertezza appropriata e che si aggiorna in risposta all'evidenza.

IV.7 — Generalizzazioni del framework: modelli gerarchici profondi e apprendimento profondo

Il framework della codifica predittiva gerarchica ha, come si è osservato nel Capitolo IV, una corrispondenza strutturale profonda con le reti neurali artificiali profonde che dominano l'attuale sviluppo dell'intelligenza artificiale. Questa corrispondenza non è puramente analogica — non si riduce a una somiglianza superficiale tra architetture — ma riflette il fatto che entrambi i tipi di sistemi sono realizzazioni di un principio computazionale comune: l'apprendimento di rappresentazioni gerarchiche che catturano la struttura causale di un ambiente attraverso la minimizzazione di un errore di predizione.

Le reti neurali profonde (DNN) con architettura encoder-decoder sono strutturalmente isomorfe, in astratto, a sistemi di codifica predittiva gerarchica. L'encoder di una DNN mappa gli input osservati su rappresentazioni latenti di ordine crescente, che corrispondono agli stati nascosti ai livelli superiori della gerarchia predittiva; il decoder mappa quelle rappresentazioni latenti su ricostruzioni degli input, che corrispondono alle predizioni generative verso il basso nella gerarchia predittiva. L'errore di ricostruzione — la differenza tra l'input originale e la sua ricostruzione — corrisponde all'errore di predizione che guida l'aggiornamento dei parametri. L'addestramento end-to-end delle DNN attraverso la discesa del gradiente sull'errore di ricostruzione è formalmente equivalente, in approssimazione, alla regola di apprendimento $\theta = -\eta \partial \mathbf{F} / \partial \theta$ del framework predittivo.

Le differenze tra le DNN standard e il framework predittivo biologicamente plausibile riguardano diversi aspetti implementativi. Le DNN standard usano algoritmo di back-propagation per

calcolare il gradiente dell'errore rispetto ai parametri — un algoritmo che richiede la trasmissione esplicita dell'informazione sull'errore da ogni livello verso i livelli inferiori, non biologicamente plausibile nella sua forma standard. Il framework predittivo biologicamente plausibile usa invece la trasmissione locale degli errori di predizione: ogni nodo riceve solo gli errori di predizione dal livello adiacente inferiore e le predizioni dal livello adiacente superiore, calcolando il gradiente localmente senza necessità di back-propagation globale. Questa struttura di apprendimento locale è biologicamente plausibile — corrispondente all'apprendimento hebbiano e alle plastiche sinaptiche locali documentate nel cervello — ed è attualmente oggetto di ricerca attiva nelle neuroscienze computazionali e nell'intelligenza artificiale, con proposte come il «target propagation» e il «predictive coding network» che esplorano come approssimare la performance della back-propagation con aggiornamenti locali.

Le reti generative variazionali (VAE) sono particolarmente prossime al framework predittivo: la VAE è esplicitamente fondata sulla minimizzazione dell'energia libera variazionale nel senso tecnico del termine, con l'encoder che implementa l'inferenza variazionale sugli stati nascosti e il decoder che implementa il modello generativo. La connessione tra le VAE e il framework predittivo frinstoniano non è solo strutturale ma formalmente precisa: l'obiettivo di addestramento della VAE — la somma del termine di ricostruzione e del termine di regolarizzazione KL — è esattamente l'energia libera variazionale nella decomposizione complessità-accuratezza presentata nell'Appendice A.2. Le VAE sono dunque, in senso preciso, implementazioni del framework predittivo variazionale nell'architettura delle reti neurali artificiali.

Questa connessione ha implicazioni per la ricerca sull'intelligenza artificiale e per la comprensione delle relazioni tra apprendimento artificiale e apprendimento biologico. Se il framework predittivo gerarchico è una buona descrizione del meccanismo dell'apprendimento percettivo e cognitivo nel cervello biologico, allora le architetture artificiali che si avvicinano maggiormente a quel framework — architetture con struttura gerarchica esplicita, inferenza variazionale, aggiornamenti locali basati sugli errori di predizione — dovrebbero generalizzare meglio a nuovi contesti e essere più efficienti nell'apprendimento rispetto ad architetture che si discostano da quella struttura. La verifica di questa previsione è uno dei programmi di ricerca più attivi al crocevia tra neuroscienze computazionali e intelligenza artificiale.

V.4 — La struttura triadica del contenuto: implicazioni per la semantica vero-condizionale

La struttura triadica del contenuto ($Q^{(l)}$, $g^{(l)}$, $\Pi^{(l)}$) ha implicazioni specifiche per la semantica vero-condizionale — la tradizione che analizza il significato in termini di condizioni di verità — che meritano di essere discusse in modo esplicito. La semantica vero-condizionale classica, da Frege a Tarski a Davidson, analizza il significato delle proposizioni come funzioni da «mondi possibili» a valori di verità: la proposizione «il gatto è sul tappeto» è vera rispetto a un mondo possibile w se e solo se in w il gatto è sul tappeto. Il contenuto semantico di una proposizione è dunque identificato con la sua funzione vero-condizionale — la funzione che mappa ogni possibile configurazione del mondo sul valore di verità della proposizione rispetto a quella configurazione.

Come si connette questa struttura vero-condizionale con la struttura triadica del contenuto nel framework predittivo? La connessione non è immediata, perché la struttura triadica è definita in termini di distribuzioni di probabilità su spazi di stati nascosti, non in termini di funzioni da mondi a valori di verità. Ma vi è una corrispondenza strutturale che si può articolare precisamente. La funzione generativa $g^{(l)}(s^{(l)})$ specifica, per ogni configurazione degli stati nascosti $s^{(l)}$, la distribuzione di probabilità sulle osservazioni attese $P(o^{(l+1)} | s^{(l)})$. Questa funzione è, formalmente, una funzione da configurazioni degli stati nascosti a distribuzioni sulle osservazioni — una funzione dal «mondo rappresentato» alle «evidenze sensoriali attese». Le condizioni di verità della rappresentazione al livello l — le condizioni sullo stato del mondo ($s^{(l)}$) rispetto alle quali la rappresentazione è corretta — sono le configurazioni degli stati nascosti su cui la distribuzione a posteriori $Q^{(l)}(s^{(l)}; \mu^{(l)})$ è concentrata: la rappresentazione è vera se e solo se lo stato del mondo effettivo è tra le configurazioni che la distribuzione a posteriori attribuisce alta probabilità.

Questa corrispondenza non è un'identità: la struttura probabilistica della distribuzione a posteriori è più ricca della struttura binaria delle condizioni di verità. Una distribuzione a posteriori concentrata su un insieme ristretto di configurazioni corrisponde a una rappresentazione con condizioni di verità precise; una distribuzione distribuita su un ampio insieme di configurazioni corrisponde a una rappresentazione con condizioni di verità vaghe o indeterminate. La vaghezza semantica — il fenomeno per cui molte espressioni del linguaggio naturale hanno condizioni di verità imprecisamente definite — trova nel framework predittivo una naturalizzazione precisa: la vaghezza è la distribuzione dell'incertezza della distribuzione a posteriori sulle configurazioni degli stati nascosti rilevanti. La frontiera tra «essere calvo» e «non essere calvo» non è nettamente definita perché la distribuzione a posteriori del sistema su «numero di capelli rilevante per la calvizie» è distribuita continuamente su un range ampio, non concentrata su un valore soglia preciso.

VI.4 — Il radicamento del contenuto morale nel framework predittivo

Una delle estensioni più provocatorie — e filosoficamente interessanti — del framework predittivo riguarda il contenuto morale: le rappresentazioni di proprietà eticamente rilevanti come il TORTO, il BENE, la GIUSTIZIA, l'OBBLIGO. La filosofia morale ha tradizionalmente resistito alle teorie naturaliste del contenuto morale, sostenendo che il contenuto morale è irriducibile alle proprietà naturali e che le proposizioni morali non sono né vere né false in senso fattuale (non-cognitivismo) o sono vere o false in virtù di fatti non naturali (non-naturalismo). Come si relaziona il framework predittivo a questo dibattito?

La posizione che si propone non è né il riduzionismo naturalistico duro — la tesi che le proprietà morali siano identiche a proprietà naturali descrivibili nei termini delle scienze empiriche — né il non-naturalismo deflazionistico — la tesi che il contenuto morale sia completamente separato dal contenuto fattuale. È piuttosto una posizione che si potrebbe chiamare *naturalismo normativo del contenuto morale*: le rappresentazioni di proprietà morali sono rappresentazioni di strutture normative che emergono dall'interazione tra sistemi cognitivi situati e ambienti sociali strutturati, rappresentazioni che hanno la struttura triadica del contenuto predittivo ma con un dominio di applicazione specifico — le relazioni tra agenti con capacità di azione e risposta affettiva.

Il contenuto del concetto TORTO, nel framework predittivo, è determinato dalla distribuzione a posteriori del sistema sulle configurazioni degli stati sociali che il sistema ha appreso a categorizzare come «torto» attraverso l'interazione con un ambiente sociale strutturato — attraverso le risposte affettive di approvazione e disapprovazione degli agenti sociali circostanti, attraverso le norme linguistiche e comportamentali del gruppo sociale di appartenenza, attraverso le proprie risposte affettive di disgusto morale, indignazione e vergogna. Questo processo di apprendimento non è diverso, strutturalmente, dal processo di apprendimento del contenuto del concetto ROSSO: entrambi coinvolgono la formazione di funzioni generative che codificano le strutture causali del dominio rilevante — il dominio delle proprietà di riflessione della luce per ROSSO, il dominio delle relazioni di azione e risposta affettiva per TORTO. La differenza è che il dominio delle proprietà morali include costitutivamente le risposte affettive e sociali degli agenti, mentre il dominio delle proprietà fisiche come il colore include proprietà fisiche che esistono indipendentemente dalle risposte degli agenti.

Questa posizione risponde al problema di Hume — del come si possa inferire un «deve» da un «è» — non riducendo il normatività morale alla fattualità descrittiva, ma mostrando che le

rappresentazioni normative sono un tipo specifico di rappresentazioni fattali: rappresentazioni delle strutture normative che caratterizzano le relazioni sociali tra agenti con capacità di azione, risposta affettiva e comunicazione. Non vi è un abisso incolmabile tra fatti e norme nel senso che il framework predittivo illumina: vi è piuttosto una continuità strutturale tra le rappresentazioni delle proprietà fisiche, le rappresentazioni delle proprietà causali, le rappresentazioni delle proprietà intenzionali degli agenti, e le rappresentazioni delle proprietà normative delle relazioni sociali. Il salto dal descrittivo al normativo non è un salto logico incolmabile ma una transizione tra livelli gerarchici del modello generativo che codificano proprietà di diverso tipo ma con la stessa struttura computazionale di base.

XI.4 — Pragmatica dell'interpretazione come inferenza predittiva

L'interpretazione pragmatica degli enunciati — il processo attraverso cui il ricevente inferisce l'intenzione comunicativa del parlante oltre il contenuto semantico letterale — è uno dei fenomeni cognitivi più complessi e stilisticamente sofisticati che l'umanità realizza con naturalezza e velocità straordinarie nella conversazione ordinaria. La normalità fenomenologica di questo processo — il fatto che i parlanti competenti non percepiscano la comprensione pragmatica come un'operazione di inferenza complessa ma come un accesso diretto al «significato» dell'enunciato — è teoricamente opaca e richiede una spiegazione che il framework predittivo può offrire in termini computazionalmente precisi.

La comprensione pragmatica richiede l'integrazione di almeno quattro livelli di informazione che il ricevente deve coordinare in tempo reale: il contenuto semantico dell'enunciato, il contesto della conversazione (incluso il contenuto degli enunciati precedenti), il contesto situazionale (la struttura della situazione in cui avviene la conversazione), e il modello degli stati mentali del parlante (credenze, desideri, intenzioni, conoscenza condivisa). L'integrazione di questi quattro livelli è computazionalmente imponente: il numero di possibili interpretazioni di un enunciato anche semplice, in un contesto realistico, è astronomicamente grande, e tuttavia i parlanti competenti convergono tipicamente sull'interpretazione intesa dal parlante in pochi centesimi di secondo.

La velocità e la robustezza della comprensione pragmatica sono spiegabili nel framework predittivo attraverso la struttura gerarchica del modello generativo della conversazione. Il ricevente non processa l'enunciato come un segnale acustico che deve essere analizzato sintatticamente, quindi semanticamente, quindi pragmaticamente in sequenza: processa l'enunciato come evidenza

per aggiornare simultaneamente distribuzioni a tutti i livelli della gerarchia predittiva — distribuzione sugli stati del mondo rilevanti per la conversazione, distribuzione sugli stati mentali del parlante, distribuzione sul significato inteso dell'enunciato. Le predizioni top-down a partire dai livelli superiori della gerarchia — le aspettative pragmatiche basate sul contesto conversazionale e situazionale — consentono al sistema di pre-attivare interpretazioni plausibili prima che l'elaborazione sensoriale bottom-up sia completata, velocizzando enormemente il processo di comprensione. La normale fluidità della comprensione conversazionale è il risultato dell'addestramento del modello generativo su migliaia di ore di conversazione che ha ottimizzato la struttura delle predizioni pragmatiche per i pattern conversazionali più frequenti nell'ambiente del parlante.

Quando la pragmatica fallisce — quando la comprensione risulta ambigua, fuorviata o erronea — il framework predittivo consente di identificare con precisione il punto del breakdown nel processo inferenziale. L'ambiguità pragmatica corrisponde a una distribuzione a posteriori che ha picchi multipli di probabilità comparabile — più interpretazioni sono ugualmente coerenti con l'evidenza disponibile. Il fraintendimento corrisponde a una distribuzione a posteriori che converge su un'interpretazione diversa da quella intesa dal parlante — a causa di prior informativi diversi da quelli del parlante, di conoscenza del contesto diversa, o di un modello degli stati mentali del parlante non calibrato sulla struttura effettiva della sua intenzione comunicativa. La difficoltà nella comprensione dell'ironia — documentata negli studi su individui con disturbi dello spettro autistico — corrisponde a una struttura di modello degli stati mentali del parlante che non include con sufficiente peso le intenzioni comunicative di secondo ordine — l'intenzione del parlante di comunicare qualcosa di diverso dal significato letterale attraverso la scelta di un enunciato letteralmente inappropriato.

Note sulla tradizione italiana di filosofia della mente

La filosofia della mente italiana ha contribuito in modo significativo al dibattito sull'intenzionalità e sul naturalismo semantico, anche se la sua produzione è meno sistematicamente citata nella letteratura anglofona rispetto alla produzione di lingua inglese, tedesca o francese. Vale la pena menzionare alcune linee della tradizione italiana che dialogano produttivamente con i temi di questa monografia.

La tradizione fenomenologica italiana — rappresentata da autori come Enzo Paci, Carlo Sini e Pier Aldo Rovatti — ha sviluppato una ricettività specifica alla fenomenologia husserliana e merleau-pontyana che condivide molte intuizioni con l'approccio del framework predittivo alla cognizione incarnata. Sini in particolare, nel suo progetto di «semiotica della pratica», ha

sviluppato una teoria del significato che enfatizza il radicamento dell'esperienza semantica nelle pratiche corporee e nei gesti — una posizione strutturalmente affine all'esternalismo semantico incarnato che il framework predittivo produce in modo computazionalmente preciso.

La tradizione analitica italiana — rappresentata da autori come Diego Marconi, Massimo Piazza e Michele Di Francesco — ha affrontato direttamente le questioni del naturalismo semantico e della competenza lessicale. Il lavoro di Marconi sulla «competenza lessicale» e la distinzione tra competenza inferenziale e competenza referenziale dei concetti linguistici è particolarmente rilevante per la discussione nel Capitolo XI sui LLM: Marconi aveva notato, prima dell'avvento dei LLM nella forma attuale, che i sistemi puramente inferenziali — sistemi che codificano solo le relazioni tra parole senza un ancoraggio alla percezione — possiedono la competenza inferenziale ma mancano della competenza referenziale. Il framework predittivo formalizza questa distinzione: la competenza inferenziale corrisponde alla struttura delle funzioni generative ai livelli superiori della gerarchia che codificano relazioni semantiche; la competenza referenziale corrisponde all'ancoraggio causale strutturale ai livelli inferiori della gerarchia che connettono i concetti alle proprietà percettive del mondo.

Considerazioni metodologiche finali

Una questione metodologica che percorre implicitamente tutta la monografia merita di essere affrontata esplicitamente in chiusura: il problema dello *status epistemico* delle affermazioni della teleosemantica computazionale sincronica. La monografia ha sviluppato argomentazioni di tipo molto diverso: argomentazioni a priori di tipo concettuale (la struttura della normatività semantica richiede funzioni proprie), argomentazioni formali (la derivazione delle equazioni di aggiornamento dall'energia libera variazionale), argomentazioni empiriche (l'evidenza neuroimaging per la struttura predittiva della corteccia), e argomentazioni pragmatiche (la dissoluzione del problema della disgiunzione). Qual è il peso relativo di questi diversi tipi di argomentazione nella valutazione complessiva della teoria?

La risposta che si propone è che la teleosemantica computazionale sincronica è una teoria che opera a un livello di astrazione intermedio tra la filosofia pura e la scienza empirica — un livello che si può caratterizzare come *epistemologia naturalizzata computazionale*. Non è pura filosofia nel senso che non si limita ad analisi concettuali a priori: le sue affermazioni hanno contenuto empirico che le rende falsificabili in linea di principio. Non è pura scienza empirica nel senso che non formula previsioni direttamente testabili sui neuroni e sui comportamenti: le sue affermazioni operano al livello computazionale — al livello degli algoritmi e delle rappresentazioni — che è intermedio tra il livello

implementazionale neurobiologico e il livello del programma concettuale filosofico.

Questa collocazione intermedia — che corrisponde alla distinzione di Marr tra livello computazionale, livello algoritmico e livello implementazionale nell'analisi dei sistemi cognitivi — ha sia vantaggi sia rischi. Il vantaggio è che le affermazioni al livello computazionale sono sia empiricamente vincolate (devono essere coerenti con l'evidenza sul comportamento cognitivo e sulla neurobiologia) sia concettualmente rigorose (devono soddisfare i requisiti di coerenza interna e di adeguatezza rispetto ai problemi filosofici che affrontano). Il rischio è che le affermazioni al livello computazionale siano più difficili da falsificare direttamente di quanto non siano le affermazioni sull'implementazione neurobiologica o i teoremi matematici: possono essere «immunizzate» rispetto all'evidenza problematica attraverso aggiustamenti al livello degli assunti accessori, rischiando di perdere il contatto critico con l'evidenza empirica che le vincola.

La consapevolezza di questo rischio è parte integrante dell'onestà intellettuale che la ricerca filosofico-scientifica richiede. Le tesi della teleosemantica computazionale sincronica sono qui presentate come proposte argomentate, non come verità stabilite: sono avanzate con l'intenzione di essere contestate, emendate e potenzialmente falsificate da sviluppi teorici ed empirici futuri. Il valore di una proposta teorica non è che sia definitivamente vera ma che articoli con precisione le domande pertinenti e identifichi le strutture rilevanti per rispondervi — che illumini il territorio del problema anche se la sua risposta specifica si rivela, in parte o in tutto, erronea. Che la teleosemantica computazionale sincronica abbia questo valore è la scommessa che questa monografia ha fatto, e la verifica di quella scommessa appartiene al futuro della ricerca.

I.6 — Externalismo semantico e internismo: la struttura del dibattito

Il dibattito tra esternalismo e internismo semantico — tra la posizione che il contenuto mentale è determinato (almeno in parte) da fattori esterni al sistema cognitivo e la posizione che è determinato esclusivamente da fattori interni — è uno dei dibattiti più fertili della filosofia della mente della seconda metà del Novecento, e le sue implicazioni si estendono ben oltre la semantica per investire questioni sull'autonomia dell'individuo cognitivo, sulla possibilità del solipsismo metodologico come strategia di ricerca, e sulla natura della conoscenza di sé.

Gli argomenti classici per l'esternalismo — l'argomento della Terra Gemella di Putnam e l'argomento dell'artrite di Burge — hanno strutture diverse che è importante distinguere. L'argomento di Putnam è un argomento sull'esternalismo *fisico*: il contenuto dei termini naturali (come «acqua», «oro», «tigre») è determinato dalla

struttura fisica del referente nel mondo reale, non dall'immagine mentale o dalla descrizione che il parlante associa al termine. Due parlanti in stati mentali intrinsecamente identici ma in ambienti con struttura fisica diversa — la Terra dove l'acqua è H_2O e la Terra Gemella dove la sostanza fenomenicamente analoga è XYZ — hanno pensieri con contenuto diverso: il parlante terrestre pensa *dell'acqua* (H_2O) quando usa il termine «acqua», il parlante della Terra Gemella pensa *dell'acqua-XYZ* (XYZ) quando usa lo stesso termine. L'argomento di Burge è un argomento sull'esternalismo *sociale*: il contenuto dei termini è determinato dalla pratica linguistica della comunità del parlante, non solo dalla struttura fisica del referente. Un parlante che crede erroneamente che l'«artrite» possa colpire i muscoli oltre alle articolazioni ha pensieri sull'artrite, non su qualche altro disturbo che colpisce anche i muscoli: il contenuto del suo pensiero è determinato dall'uso del termine nella sua comunità linguistica, non dalla sua comprensione individuale incompleta.

Questi due argomenti, pur convergendo sulla conclusione che il contenuto è esternamente determinato, hanno fonti molto diverse: l'uno radica il contenuto nella struttura fisica del mondo naturale, l'altro nella pratica linguistica del gruppo sociale. La combinazione delle due forme di esternalismo — il naturalismo dei tipi naturali e il socialismo dei tipi sociali — produce una posizione complessa che è difficile da unificare in modo coerente: le due fonti di determinazione del contenuto possono in linea di principio entrare in conflitto, come nei casi in cui la pratica linguistica di una comunità designa con un termine una proprietà diversa da quella che l'uso scientifico del termine designa.

Il framework della teleosemantica computazionale sincronica incorpora entrambe le forme di esternalismo in modo coerente attraverso la struttura del modello generativo. L'esternalismo fisico putnamiano è incorporato nell'ancoraggio causale strutturale: il modello generativo del sistema ha la struttura che ha perché si è formato attraverso l'interazione causale con l'ambiente fisico, e le funzioni generative ai livelli superiori della gerarchia codificano le strutture causali reali dell'ambiente — incluse le strutture dei tipi naturali come l'acqua, l'oro, i metalli — piuttosto che le proprietà superficiali che i parlanti ignoranti della struttura chimica profonda associano a quei tipi. L'esternalismo sociale burgeano è incorporato nella dipendenza del modello generativo dai segnali comunicativi degli altri agenti: il modello linguistico che il sistema acquisisce dipende dagli usi linguistici della comunità, che mediano l'accesso alla struttura semantica del linguaggio pubblico e dei suoi termini.

La coerenza tra le due forme di esternalismo emerge dalla struttura gerarchica del modello generativo: i livelli superiori della gerarchia codificano strutture causali profonde (la struttura chimica dell'acqua, le proprietà reali dell'artrite) che sono apprese attraverso la mediazione sociale (la pratica linguistica della

comunità che usa il termine in modo calibrato sulla struttura reale del referente). I due livelli di determinazione del contenuto non entrano in conflitto perché operano a livelli gerarchici diversi: il livello delle strutture causali profonde e il livello delle pratiche linguistiche sociali sono entrambe componenti del modello generativo complessivo, con le pratiche sociali che forniscono l'accesso mediato alle strutture causali che i parlanti individualmente non possono osservare direttamente.

II.5 — Teleosemantica e anti-realismo: il problema delle funzioni nelle specie estinte

Un aspetto della teleosemantica millikaniana che questa monografia non ha discusso sufficientemente è il problema che si pone per le funzioni proprie degli organismi che non hanno discendenza — le specie estinte. Se la funzione propria di un organo è determinata dalla sua storia di selezione, quale è la funzione propria del cuore di un dinosauro vissuto e morto senza discendenza? Il cuore di quel dinosauro ha pompato sangue; quell'effetto era la ragione della sopravvivenza degli antenati del dinosauro che avevano cuori; ma il dinosauro in questione non ha avuto discendenza, per cui il suo cuore non ha contribuito direttamente alla perpetuazione della linea. Ha ancora una funzione propria?

Millikan risponde che la funzione propria di un organo è determinata dalla storia della linea genealogica della specie, non dalla riproduzione del singolo individuo: l'organo di un individuo senza discendenza ha una funzione propria in virtù della storia della sua linea genealogica fino a quel punto, anche se quella linea termina con lui. Questa risposta è plausibile ma solleva una domanda successiva: come si estende la funzione propria a strutture che non fanno parte di una linea genealogica — strutture artificiali, organismi creati in laboratorio, sistemi cognitivi artificiali? Se la funzione propria richiede una linea genealogica, i sistemi artificiali non hanno funzioni proprie in senso stretto, il che implica che non hanno contenuto semantico — una conclusione difficile da accettare.

La teleosemantica computazionale sincronica risponde a questo problema in modo diretto, perché la normatività semantica è fondata sulla struttura sincronica del modello generativo piuttosto che sulla storia genealogica. Un sistema artificiale che ha acquisito un modello generativo attraverso l'addestramento su dati ambientali ha funzioni proprie nel senso rilevante per la normatività semantica: le sue funzioni proprie sono determinate dalla struttura del modello generativo e dalla storia di addestramento che lo ha prodotto, non da una linea genealogica biologica. La «storia selettiva» rilevante è la storia dell'ottimizzazione — qualsiasi processo che ha selezionato le

configurazioni del modello che producono previsioni accurate — non necessariamente la selezione biologica darwiniana. Questa generalizzazione consente di estendere la teoria del contenuto ai sistemi artificiali senza richiedere che abbiano una storia evolutiva biologica.

VIII.8 — La coscienza negli animali non umani: implicazioni del framework predittivo

La questione della coscienza animale — in che misura e in quale forma gli animali non umani hanno esperienza soggettiva — è una delle questioni scientifiche ed etiche più urgenti che il framework predittivo può illuminare con strumenti teorici precisi. La difficoltà tradizionale nell'affrontare questa questione è il problema metodologico dell'accesso in prima persona: non possiamo sapere direttamente se un animale ha esperienza soggettiva, e le evidenze comportamentali sono notoriamente ambigue — possono essere prodotte sia da sistemi con esperienza soggettiva sia da sistemi «zombie» funzionalmente equivalenti senza esperienza.

Il framework predittivo offre un approccio indiretto ma computazionalmente preciso: invece di cercare evidenza diretta dell'esperienza soggettiva, si può caratterizzare la complessità del modello generativo di un organismo — la ricchezza delle sue rappresentazioni gerarchiche, la sofisticazione del suo modello di sé, la complessità delle sue strutture di precisione — e usare queste misure come proxy della ricchezza della sua esperienza. Non si pretende di risolvere il problema difficile attraverso questo approccio: si assume solo che vi sia una correlazione sistematica tra la complessità computazionale del modello generativo e la ricchezza dell'esperienza soggettiva, correlazione che è plausibile ma non logicamente necessaria.

Sulla base di questa correlazione, si può articolare una scala graduata della coscienza animale nel framework predittivo. Gli organismi unicellulari — batteri, amebe — hanno comportamento chemiotattico adattivo che può essere descritto in termini di minimizzazione dell'energia libera a livello cellulare, ma la semplicità del «modello» che implementano — essenzialmente una risposta diretta ai gradienti chimici senza rappresentazioni gerarchiche — suggerisce che, se hanno esperienza, questa è di tipo elementarissimo. Gli invertebrati con sistemi nervosi semplici — nematodi, insetti con cervelli di poche centinaia di neuroni — hanno modelli generativi leggermente più complessi che consentono forme semplici di apprendimento e comportamento contestuale, ma la mancanza di struttura gerarchica elaborata suggerisce che le loro esperienze, se presenti, siano di tipo molto diverso dall'esperienza umana. I vertebrati a sangue caldo — uccelli e mammiferi — hanno strutture neurali con gerarchia corticale o equivalente funzionale (nei volatili, la pallium) che

consentono modelli generativi gerarchicamente strutturati, modelli di sé più complessi, e strutture di precisione neuromodulatoria. L'evidenza comportamentale suggerisce che uccelli corvidi e grandi scimmie abbiano forme di autoconsapevolezza (test dello specchio), empatia e pianificazione del futuro coerenti con modelli generativi di sé sufficientemente ricchi da supportare una forma di esperienza soggettiva qualitativamente rilevante, anche se diversa dall'esperienza umana.

Le implicazioni etiche di questa graduazione sono significative ma non univoche. Se la ricchezza dell'esperienza è proporzionale alla complessità del modello generativo, allora le preoccupazioni morali per il benessere animale dovrebbero essere graduatamente scalate con la complessità del sistema nervoso — una conclusione che è già implicita in molte tradizioni etiche ma che il framework predittivo sostiene con basi teoriche più precise. La questione difficile è dove tracciare la soglia della considerazione morale — a quale livello di complessità del modello generativo emerge la sofferenza moralmente rilevante — e questa soglia non è determinata dalla teoria computazionale ma da scelte normative aggiuntive che il framework non può determinare autonomamente.

XII — Prospettive di ricerca futura

Sebbene la monografia non sia stata progettata con un capitolo conclusivo numerato XII, è opportuno identificare in modo sistematico le principali direzioni di ricerca aperte dalla proposta teorica sviluppata, per orientare i lettori interessati a contribuire all'avanzamento del programma di ricerca che essa delinea.

La prima direzione è la *formalizzazione computazionale del contenuto astratto*. Il Capitolo VI ha sviluppato una trattazione dei concetti astratti come CAUSALITÀ, NUMERO e GIUSTIZIA che ha carattere prevalentemente qualitativo: ha indicato la struttura generale del radicamento del contenuto astratto nel modello generativo gerarchico ma non ha fornito un'implementazione computazionale dettagliata. La ricerca futura dovrebbe affrontare la questione di come le funzioni generative ai livelli superiori della gerarchia possano implementare strutture sufficientemente ricche da catturare la composizionalità e la sistematicità richieste dal pensiero astratto, e di come tali strutture possano essere apprese da sistemi artificiali o biologici attraverso la minimizzazione dell'energia libera.

La seconda direzione è la *integrazione con la semantica formale del linguaggio naturale*. La semantica formale sviluppata nella tradizione Montague-Lewis fornisce un framework preciso per la struttura vero-condizionale del linguaggio naturale — inclusa la semantica dei quantificatori, dei modali, dei condizionali e delle strutture intensionali — che il framework predittivo non ha ancora adeguatamente incorporato. La connessione tra le distribuzioni di

probabilità del framework predittivo e le funzioni intensionali della semantica formale — entrambe funzioni da mondi possibili a valori — suggerisce che una tale integrazione sia in linea di principio realizzabile, ma le difficoltà tecniche sono considerevoli e richiedono lavoro specifico all'intersezione tra semantica formale e scienza cognitiva computazionale.

La terza direzione è la *verifica empirica sistematica delle previsioni neuroscientifiche*. Le previsioni articolate nel Capitolo IX sulla struttura gerarchica della specializzazione semantica, sul ruolo della precisione nell'elaborazione semantica e sulla struttura degli errori semantici in condizioni di carico cognitivo non sono ancora state testate sistematicamente come previsioni specifiche della teleosemantica computazionale sincronica. Un programma di ricerca sperimentale che le verifichi metodicamente — utilizzando tecniche di neuroimaging ad alta risoluzione, di elettrofisiologia multiunitaria e di manipolazione farmacologica — fornirebbe evidenza cruciale per la valutazione empirica della teoria.

La quarta direzione è l'*analisi comparativa con altre teorie naturaliste del contenuto*. La monografia ha discusso le relazioni con la teleosemantica millikaniana, la semantica informazionale dretsiana e il fodorismo, ma non ha confrontato sistematicamente la proposta con alcune teorie più recenti che meriterebbero attenzione: la teoria dell'implicatura bayesiana di Goodman e Tenenbaum, il framework del contenuto come copula di Brown, la teoria della categoria ontologica di Lowe, e la proposta neomillikaniana di Schulte. Un confronto sistematico con queste teorie illuminerebbe le differenze specifiche e i vantaggi comparativi della teleosemantica computazionale sincronica rispetto alle alternative disponibili.

La quinta direzione è lo *sviluppo di architetture artificiali ispirate al framework predittivo biologicamente plausibile* che realizzino le strutture di inferenza attiva, aggiornamento locale degli errori di predizione e modulazione della precisione in modo che superi le limitazioni delle architetture correnti. Questi sviluppi tecnici non sono di interesse puramente ingegneristico: fornirebbero anche test computazionali delle previsioni teoriche sul tipo di contenuto semantico che diversi tipi di architetture predittive possono realizzare, permettendo una forma di sperimentazione computazionale che completerebbe l'evidenza neuroscientifica empirica.

L'ampiezza e la varietà di queste direzioni di ricerca conferma che la teleosemantica computazionale sincronica è un programma di ricerca in progress, non una teoria compiuta. La sua forza non è la completezza ma la coerenza: le tesi nucleari presentate nel Capitolo IX forniscono un nucleo strutturale sufficientemente robusto da orientare la ricerca in tutte queste direzioni, generando previsioni testabili e identificando le questioni aperte che richiedono lavoro futuro. Un programma di ricerca che generi previsioni falsificabili e identifichi le proprie lacune è, secondo il criterio popperiano, più

scientifico di una teoria che si presenta come completa e non richiede sviluppi ulteriori. La teleosemantica computazionale sincronica, in questo senso, aspira a essere scientificamente virtuosa proprio nella misura in cui è esplicitamente incompleta.

Note bibliografiche

Avvertenza sul metodo bibliografico

La presente monografia adotta un metodo bibliografico che privilegia i testi primari — i contributi originali degli autori citati nelle loro formulazioni più autorevoli — rispetto ai commenti e alle rassegne secondarie. I riferimenti bibliografici hanno funzione di ancoraggio argomentativo: ogni riferimento indica la fonte specifica dell'argomentazione o della tesi attribuita all'autore, consentendo al lettore di verificare la fedeltà dell'attribuzione e di approfondire autonomamente il contesto teorico originale.

Riferimenti per capitolo

Capitolo I. La tesi dell'intenzionalità come Grundcharakter della coscienza è sviluppata in Brentano (1874/2009). La distinzione senso-denotazione è introdotta da Frege (1892/2007). La teoria delle descrizioni è in Russell (1905). La critica del descrittivismo e la teoria causale della riferimento sono in Kripke (1972/1980). L'argomento della Terra Gemella è sviluppato in Putnam (1975). L'esternalismo sociale è in Burge (1979). La critica dell'IA come mera manipolazione sintattica è in Searle (1980, 1983). Il principio del triangolo interpretativo è in Davidson (1984). Il problema della causalità mentale è sistematizzato in Kim (1993, 2005).

Capitolo II. La teoria delle funzioni proprie e del BIVC è in Millikan (1984, 1993). La formulazione alternativa teleosemantica è in Papineau (1987). La versione eziologica della teleosemantica è in Neander (1991a, 1991b). La critica delle funzioni proprie intese diacronicamente è in Boghossian (1989).

Capitolo III. La teoria dell'informazione come riduzione dell'incertezza è in Shannon e Weaver (1949). La naturalizzazione del contenuto informazionale è in Dretske (1981). La proposta delle condizioni periferiche per risolvere la disgiunzione è in Dretske (1988). La critica del problema della disgiunzione è in Fodor (1987, 1990).

Capitolo IV. Il principio della minimizzazione dell'energia libera come principio unificante della cognizione biologica è elaborato in Friston (2010, 2013). La connessione con l'inferenza bayesiana cerebrale è in Knill e Pouget (2004) e Weiss, Simoncelli e Adelson (2002). La formulazione in termini di codifica predittiva gerarchica è in Rao e Ballard (1999) e Clark (2016). Il legame con il calcolo variazionale e l'approssimazione di Laplace è in Friston, Trujillo-Barreto e Daunizeau (2008).

Capitolo V. La dissoluzione del problema della disgiunzione attraverso la struttura triadica è sviluppata originalmente in questa monografia in dialogo con Millikan (1984) e Fodor (1987, 1990). Il problema della normatività nel contesto predittivo è affrontato in Wiese (2018) e Gładziejewski (2016). Il contenuto percettivo nel framework predittivo è analizzato in Friston (2003, 2005).

Capitolo VI. Il radicamento del contenuto astratto nel framework predittivo è sviluppato in dialogo con Barsalou (1999), Lawrence e Caramazza (2011) e Pulvermüller (2013). I concetti incarnati e la cognizione situated sono discussi in Wilson (2002) e Shapiro (2011). Il caso del concetto di causalità nel framework predittivo è in Tenenbaum, Kemp, Griffiths e Goodman (2011).

Capitolo VII. La critica searleana all'IA e al computazionalismo è in Searle (1980, 1992). La risposta anti-regressiva è sviluppata in dialogo con Dennett (1987) e Dretske (1988). La distinzione tra ancori causali strutturali e attribuzione convenzionale è originale di questa monografia.

Capitolo VIII. Il problema difficile della coscienza è introdotto in Chalmers (1995, 1996). Il sé minimale e il sé narrativo sono discussi in Zahavi (2005). Il legame tra sistemi neuromodulatori e precisione è sviluppato in Friston et al. (2012). La coscienza come inferenza predittiva è in Hohwy (2013) e Clark (2019). Gli psichedelici nel framework predittivo sono analizzati in Carhart-Harris e Friston (2019).

Capitolo IX. La teoria dello spazio di lavoro globale è in Baars (1988) e Dehaene (2014). La teoria dell'informazione integrata è in Tononi (2004, 2008). La causalità mentale nel contesto predittivo è discussa in Friston, Wiese e Metzinger (2021).

Capitolo X. La teoria del linguaggio del pensiero è in Fodor (1975, 2008). La teoria allostatica del contenuto affettivo è in Corcoran, Pezzulo e Friston (2020) e Bhatt, Bhatt e Bhatt (2020, in dialogo con Friston 2010). La naturalizzazione delle condizioni di verità è discussa in Field (1972, 1978).

Capitolo XI. L'evidenza neuroimaging per la codifica predittiva è in Bastos et al. (2012) e Clark (2016). La componente N400 come misura della predizione semantica è in Kutas e Federmeier (2011). Il modello RSA è in Frank e Goodman (2012) e Goodman e Frank (2016). I LLM come sistemi predittivi sono discussi in Linzen e Baroni (2021) e McClelland et al. (2020).

Bibliografia completa

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4), 577–660.

- Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4), 695–711.
- Boghossian, P. A. (1989). The rule-following considerations. *Mind*, 98(392), 507–549.
- Brentano, F. (1874/2009). *Psicologia dal punto di vista empirico* (trad. it.). Laterza. (ed. orig. 1874).
- Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, 4(1), 73–121.
- Carhart-Harris, R. L., & Friston, K. J. (2019). REBUS and the anarchic brain: Toward a unified model of the brain action of psychedelics. *Pharmacological Reviews*, 71(3), 316–344.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
- Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.
- Clark, A. (2019). Consciousness as generative entanglement. *Journal of Philosophy*, 116(12), 645–662.
- Corcoran, A. W., Pezzulo, G., & Friston, K. J. (2020). From allostatic agents to counterfactual cognisers: Active inference, biological regulation, and the origins of cognition. *Biology & Philosophy*, 35(3), 1–45.
- Davidson, D. (1984). *Inquiries into truth and interpretation*. Oxford University Press.
- Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Viking.
- Dennett, D. C. (1987). *The intentional stance*. MIT Press.
- Dretske, F. (1981). *Knowledge and the flow of information*. MIT Press.
- Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. MIT Press.
- Field, H. (1972). Tarski's theory of truth. *Journal of Philosophy*, 69(13), 347–375.
- Field, H. (1978). Mental representation. *Erkenntnis*, 13(1), 9–61.
- Fodor, J. A. (1975). *The language of thought*. Harvard University Press.
- Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.
- Fodor, J. A. (1990). *A theory of content and other essays*. MIT Press.

- Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Oxford University Press.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998–998.
- Frege, G. (1892/2007). Senso e denotazione. In *Logica e aritmetica* (trad. it., pp. 372–403). Boringhieri. (ed. orig. 1892).
- Friston, K. J. (2003). Learning and inference in the brain. *Neural Networks*, 16(9), 1325–1352.
- Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360(1456), 815–836.
- Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.
- Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, 10(86), 20130475.
- Friston, K. J., Trujillo-Barreto, N., & Daunizeau, J. (2008). DEM: A variational treatment of dynamic systems. *NeuroImage*, 41(3), 849–885.
- Friston, K. J., Wiese, W., & Metzinger, T. (2021). Minimizing surprise and the phenomenology of consciousness. *Frontiers in Psychology*, 12, 698831.
- Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, 193(2), 559–582.
- Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11), 818–829.
- Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
- Kim, J. (1993). *Supervenience and mind: Selected philosophical essays*. Cambridge University Press.
- Kim, J. (2005). *Physicalism, or something near enough*. Princeton University Press.
- Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12), 712–719.
- Kripke, S. A. (1972/1980). *Naming and necessity*. Harvard University Press.
- Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62, 621–647.
- Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, 7, 195–212.

- McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42), 25966–25974.
- Merleau-Ponty, M. (1945/2003). *Fenomenologia della percezione* (trad. it.). Bompiani. (ed. orig. 1945).
- Millikan, R. G. (1984). *Language, thought, and other biological categories*. MIT Press.
- Millikan, R. G. (1993). *White queen psychology and other essays for Alice*. MIT Press.
- Neander, K. (1991a). Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science*, 58(2), 168–184.
- Neander, K. (1991b). The teleological notion of 'function'. *Australasian Journal of Philosophy*, 69(4), 454–468.
- Papineau, D. (1987). *Reality and representation*. Blackwell.
- Pulvermüller, F. (2013). How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences*, 17(9), 458–470.
- Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7, 131–193.
- Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87.
- Russell, B. (1905). On denoting. *Mind*, 14(56), 479–493.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
- Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.
- Searle, J. R. (1992). *The rediscovery of the mind*. MIT Press.
- Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.
- Shapiro, L. (2011). *Embodied cognition*. Routledge.
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285.
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5(1), 42.

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215(3), 216–242.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6), 598–604.

Wiese, W. (2018). Toward a mature science of consciousness. *Frontiers in Psychology*, 9, 693.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4), 625–636.

Zahavi, D. (2005). *Subjectivity and selfhood: Investigating the first-person perspective*. MIT Press. # ESPANSIONI — Sezioni aggiuntive e approfondimenti

Supplemento bibliografico annotato

La presente sezione integra la bibliografia completa con note di orientamento critico destinate al lettore che desideri approfondire i temi principali trattati nella monografia. Le annotazioni non sostituiscono la lettura dei testi primari ma indicano la rilevanza specifica di ciascun testo per le argomentazioni sviluppate e segnalano le principali obiezioni e discussioni che hanno seguito la loro pubblicazione.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? Questo articolo è il documento fondativo del framework dell'inferenza attiva nella sua forma matura. Friston presenta il principio di minimizzazione dell'energia libera come un principio unificante che spiega l'intera gamma del comportamento biologico adattivo — dalla percezione all'azione, dall'attenzione all'apprendimento, dalla cognizione alle emozioni — nei termini di un unico processo di minimizzazione variazionale. La presentazione è accessibile ma tecnicamente rigorosa, e l'articolo è indispensabile per comprendere le fondamenta matematiche del framework sviluppato nei Capitoli IV e seguenti. Una lettura critica dovrebbe tenere presente l'ambizione sistematica del framework, che ha generato riserve da parte di chi ritiene che la sua generalità lo renda difficilmente falsificabile: se ogni comportamento biologico è «minimizzazione dell'energia libera», la teoria rischia di diventare un'etichetta descrittiva piuttosto che un'ipotesi genuinamente testabile.

Millikan, R. G. (1984). Language, Thought, and Other Biological Categories. Opera fondativa della teleosemantica evolutivista. Il capitolo sulle funzioni proprie e sul principio del BIVC (Biological Information Very Cheaply) è il nucleo teorico da cui discende l'intera tradizione teleosemantica. La monografia di Millikan è scritta in uno stile filosofico denso e tecnico che richiede pazienza al lettore non specializzato, ma il sistema teorico che

costruisce è tra i più rigorosi e coerenti disponibili nella filosofia del linguaggio naturalistica. Centrale per i Capitoli II e V.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Presentazione accessibile e sistematicamente organizzata del framework predittivo con enfasi sull'azione incorporata e sul ruolo dell'agentività nell'inferenza attiva. Clark è più attento alla tradizione fenomenologica e all'integrazione con le neuroscienze cognitive rispetto a Friston, e il libro è ideale per il lettore che viene dalla filosofia della mente e vuole comprendere le implicazioni del framework per la teoria dell'incarnazione. Il dibattito sulla «mente predittiva» di Clark ha ricevuto risposte di grande interesse da Hohwy, Jakob Hohwy, e Andy Clark stesso, che sono raccolte in una serie di scambi nel *Journal of Consciousness Studies* e in *Consciousness and Cognition*.

Hohwy, J. (2013). *The Predictive Mind*. Il testo filosoficamente più rigoroso e sistematicamente sviluppato sul framework predittivo in relazione alla filosofia della mente. Hohwy è attento alle questioni di demarcazione — cosa distingue davvero il framework predittivo dalle sue alternative — e ai problemi di vaghezza e circolarità che alcune formulazioni del principio di energia libera rischiano di incorporare. Il capitolo sulla percezione come inferenza inconscia e quello sulla relazione tra il framework e il problema difficile della coscienza sono particolarmente rilevanti per i Capitoli V-VIII di questa monografia.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Il testo fondativo della semantica informazionale naturalistica. La prima parte — sulla teoria dell'informazione come base per il contenuto cognitivo — è tecnicamente esigente ma fondamentale; la seconda parte sulla conoscenza e l'epistemia è meno rilevante per i temi di questa monografia. La critica di Fodor al libro di Dretske in *Psychosemantics* è un buon punto di partenza per comprendere le obiezioni al programma informazionale. Centrale per il Capitolo III.

Chalmers, D. J. (1996). *The Conscious Mind*. La formulazione più sistematica del problema difficile della coscienza e dell'argomento degli zombie. Chalmers sviluppa in modo rigoroso la tesi dell'irriducibilità dell'esperienza soggettiva al funzionale, con implicazioni per qualsiasi teoria naturalista del contenuto che pretenda di ridurre il contenuto fenomenico al contenuto funzionale. La risposta del Capitolo VIII di questa monografia — la distinzione tra i correlati funzionali della coscienza e il problema difficile residuo — è elaborata in dialogo con la posizione di Chalmers.

Zahavi, D. (2005). *Subjectivity and Selfhood: Investigating the First-Person Perspective*. Il testo più influente sulla fenomenologia del sé nella tradizione contemporanea che dialoga con la scienza cognitiva. Zahavi sviluppa la distinzione tra sé

minimale e sé narrativo — rilevante per il Capitolo VIII — e difende l'importanza della prospettiva in prima persona contro le teorie eliminativiste del sé. Il dialogo tra Zahavi e i difensori del framework predittivo sulla natura del sé è raccolto in diversi scambi in *Phenomenology and the Cognitive Sciences*.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. Articolo fondativo del modello RSA (Rational Speech Act) nella sua versione più matura. L'articolo presenta la formalizzazione bayesiana dell'interpretazione pragmatica con applicazioni alla semantica scalare, ai riferimenti visivi e alla comprensione delle metafore. Rilevante per la sezione XI.5 e per la connessione tra RSA e il framework predittivo sviluppata in quel capitolo.

Kim, J. (1993). Supervenience and Mind. Raccolta degli articoli fondamentali di Kim sul problema della causazione mentale e sulla riduzione delle proprietà mentali. Il capitolo sul problema dell'esclusione è il riferimento fondamentale per la discussione del Capitolo IX.2 sulle implicazioni della teleosemantica computazionale sincronica per la causazione mentale. La posizione di Kim ha generato un dibattito vastissimo; tra le risposte più interessanti si segnalano quelle di Yablo, Block, McLaughlin e Fodor.

Brentano, F. (1874/2009). Psicologia dal punto di vista empirico. Il testo in cui Brentano reintroduce la nozione di intenzionalità — l'inesistenza intenzionale — come Grundcharakter del mentale. Il primo capitolo del Libro II è il luogo di partenza di tutta la tradizione analitica e fenomenologica sull'intenzionalità. La traduzione italiana di Borsari per Laterza è accurata e include una buona introduzione critica.

Indice dei concetti principali

Per orientare il lettore nella struttura argomentativa della monografia, si offre qui un indice sintetico dei concetti principali con indicazione dei capitoli in cui vengono introdotti, sviluppati e criticati.

Ancoraggio causale strutturale — introdotto nel Capitolo VII come risposta alla critica searleana dell'intenzionalità derivata; sviluppato nelle sezioni VII.2 e VII.3; connesso alla tesi II della teoria nel Capitolo IX.

Approssimazione di Laplace — introdotta nel Capitolo IV come tecnica variazionale; formalizzata nell'Appendice A.3; connessa all'inferenza bayesiana esatta nel limite di alta precisione.

Complessità-accuratezza — decomposizione dell'energia libera variazionale introdotta nel Capitolo IV e sviluppata nell'Appendice

A.2; centrale per la comprensione del bilanciamento tra semplicità del modello e adattamento ai dati.

Contenuto astratto — il problema centrale del Capitolo VI; rilevante per la critica della teleosemantica diacronica e per la difesa della svolta sincronica; connesso alla tesi VI della teoria.

Energia libera variazionale — introdotta nel Capitolo IV; formalizzata nell'Appendice A.2; il concetto matematico centrale dell'intero framework.

Funzione propria sincronica — il concetto che distingue la teleosemantica computazionale sincronica dalla teleosemantica millikaniana; introdotto nel Capitolo V; fondamento della tesi V; discusso criticamente nell'Appendice C.4.

Incarnazione — trattata nel Capitolo VII in relazione all'intenzionalità originaria; approfondita nell'Appendice C.5 in relazione al corpo proprio e all'inferenza attiva.

Inferenza attiva — estensione del framework predittivo all'azione; introdotta nel Capitolo VII; formalizzata nell'Appendice A.6; connessa alla tesi del sé come agente inferenziale attivo.

Modello di sé — il concetto centrale del Capitolo VIII; definito formalmente come la componente del modello generativo che rappresenta l'agente stesso; connesso alla distinzione tra sé minimale e sé narrativo.

Precisione — parametro formale nel framework predittivo che determina il peso degli errori di predizione; introdotto nel Capitolo IV; connesso all'attenzione selettiva e ai sistemi neuromodulatori nel Capitolo VIII; formalizzato nell'Appendice A.7.

Problema della disgiunzione — introdotto nel Capitolo III in relazione alla semantica dretsiana; la sua dissoluzione nel framework predittivo è sviluppata nel Capitolo V e nella tesi III della teoria nel Capitolo IX; analizzata formalmente nella sezione III.3.

Struttura triadica del contenuto — il concetto centrale della proposta teorica; introdotto nel Capitolo V; fondamento della tesi I; sviluppato nelle implicazioni per la semantica vero-condizionale nella sezione V.4.

Teleosemantica computazionale sincronica — il nome della teoria proposta; introdotto nel Capitolo IX; le sei tesi nucleari sono formulate nel paragrafo IX.1; le implicazioni per la causazione mentale, le previsioni neuroscientifiche e il confronto con GWT e IIT sono svolte nei paragrafi successivi.

Nota sull'uso del formalismo matematico

Una scelta metodologica che richiede giustificazione esplicita è l'uso estensivo del formalismo matematico in una monografia di filosofia. La tradizione analitica ha a lungo discusso se e in che misura il formalismo matematico sia appropriato nella filosofia della mente e nella semantica: da un lato, il formalismo può chiarire strutture teoriche che il linguaggio naturale esprime ambigualmente; dall'altro, può dare un'apparenza di precisione a costrutti teorici che la matematica non rende genuinamente più chiari ma semplicemente diversamente opachi.

La posizione metodologica adottata in questa monografia è che il formalismo matematico è giustificato quando e nella misura in cui il framework teorico che si formalizza è genuinamente matematico — quando la struttura del framework dipende da relazioni quantitative tra grandezze che la matematica consente di specificare con precisione che il linguaggio naturale non può raggiungere. Il framework della codifica predittiva gerarchica è genuinamente matematico in questo senso: la distinzione tra errore di predizione e errore di predizione ponderato dalla precisione non è una distinzione che si può fare chiaramente nel linguaggio naturale senza il formalismo delle matrici di precisione; la derivazione dell'energia libera variazionale come limite superiore della sorpresa non è chiara senza l'algebra della divergenza KL; la connessione tra il filtro di Kalman e la codifica predittiva non è visibile senza la linearizzazione formale. In questi casi, il formalismo non è ornamento: è il veicolo attraverso cui le tesi filosofiche ricevono la specificazione precisa che le rende testabili.

D'altro canto, il formalismo è stato deliberatamente limitato nei punti in cui la matematica sarebbe stata un'apparenza di precisione senza sostanza: le discussioni della fenomenologia husserliana, delle critiche searleane e delle implicazioni per la causazione mentale e la coscienza sono condotte nel linguaggio naturale tecnico della filosofia, perché il tipo di distinzioni che quelle discussioni richiedono è di tipo concettuale e non quantitativo. La monografia aspira a usare il formalismo come strumento del pensiero, non come sostituto del pensiero.

Le equazioni presentate sono in notazione Unicode che consente la lettura diretta senza software specializzato: le lettere greche, i simboli matematici e le notazioni vettoriali sono scritti in caratteri Unicode standard. Questa scelta ha un costo in leggibilità rispetto alla notazione LaTeX, ma ha il vantaggio di rendere il testo accessibile in qualsiasi ambiente di visualizzazione senza dipendenze software. Il lettore che desidera trasferire le equazioni in un sistema di composizione tipografica LaTeX per uso accademico troverà la conversione meccanicamente semplice.

Note sulla traduzione terminologica dall'inglese

La letteratura sul framework predittivo e sull'inferenza attiva è prevalentemente in inglese, e i termini tecnici di quella letteratura non hanno ancora una traduzione italiana stabilizzata. Le scelte traduttive adottate in questa monografia sono le seguenti, con la motivazione delle scelte nei casi in cui queste si discostano da proposte alternative disponibili nella letteratura italiana.

Free energy — tradotto come «energia libera» nel senso della termodinamica statistica da cui il termine è importato per analogia. Alcune traduzioni parlano di «energia di sorpresa» (dal termine «surprise» con cui $-\log P(\theta)$ è talvolta indicato) o di «energia variazionale», ma «energia libera variazionale» è il termine più fedele alla struttura matematica del concetto e alla sua derivazione dalla termodinamica di Helmholtz.

Predictive coding — tradotto come «codifica predittiva» seguendo la maggioranza delle trattazioni in italiano disponibili. Alternative come «codificazione predittiva» o «codice predittivo» sono meno fluide e non hanno trovato largo impiego.

Active inference — tradotto come «inferenza attiva» in modo letterale. Il termine esprime bene il concetto: l'inferenza che include l'azione come componente, non solo la formazione di credenze passive sull'ambiente.

Prediction error — tradotto come «errore di predizione» in modo diretto. L'alternativa «errore di previsione» ha connotazioni che invitano a confusioni con la previsione in senso ordinario; «predizione» mantiene il senso tecnico della previsione del sistema generativo sulle osservazioni attese.

Precision — tradotto come «precisione» nel senso tecnico della matrice di precisione gaussiana — l'inversa della covarianza. In italiano, «precisione» ha questo significato tecnico in statistica e probabilità, e l'uso nella codifica predittiva è coerente con l'uso statistico standard.

Generative model — tradotto come «modello generativo», seguendo l'uso consolidato in italiano nella letteratura sull'apprendimento automatico e sulla scienza cognitiva computazionale.

Proper function — tradotto come «funzione propria» seguendo la traduzione consolidata in italiano della terminologia millikaniana. L'alternativa «funzione appropriata» cattura meno bene il senso di «funzione che un sistema ha per il fatto di essere il tipo di sistema che è» che Millikan intende.